ARI Research Note 88-36

IMPROVING THE SELECTION CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL: SU[illegible]

Annual Report, 1986 Fiscal Year
Supplement to
ARI Technical Report 792

John P. Campbell, Editor
Human Resources Research Organization

for

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

MANPOWER AND PERSONNEL RESEARCH LABORATORY
Newell K. Eaton, Director

U. S. Army
Research Institute for the Behavioral and Social Sciences

May 1988

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

	L. NEALE COSBY
EDGAR M. JOHNSON	Colonel, IN
Technical Director	Commander

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Mark Czarnoleski

This report, as submitted by the contractor, has been cleared for release to Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or other reference services such as the National Technical Information Service (NTIS). The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

ARI RESEARCH NOTE 88-36

16. Supplementary Notation (continued)

American Institutes for Research, Personnel Decisions Research Institute, and Army Research Institute

18. Subject Terms (continued)

Soldier Effectiveness

EDITORS' PREFACE

In the course of executing the mainline research program of Project A, it has always been an accepted--indeed priority--practice to find mechanisms and means for communicating and sharing early and/or otherwise salient research results and activities with the U.S. Army and with the professional research community at large. As a result, numerous papers, reports, and symposium proceedings have been produced each year to meet the continuing interest of both scientific and operational audiences. The custom within Project A has been to compile these documents and to publish them as an adjunct to the Project A Annual Report.

The reports in this Supplement to the Fiscal Year 1986 Annual Report are presented in chronological order. Most of them are referenced in the Annual Report. That some are not should in no way diminish their importance or relevance to the readers of these reports. Each document was produced to meet a specific need and audience and, when taken in context, provides, in effect, a chronology of reports and communications which can reveal the process and flow of the overall research program being accomplished collegially by the U.S. Army Research Institute and contractor scientists. In many cases these findings have been further refined or synthesized into more formal contract-deliverable items.

Lawrence M. Hanser

Lola M. Zook

CONTENTS

Page

Utility Estimation in Five Enlisted Occupations. 1

Newell K. Eaton, Hilda Wing, and Alan Lau (ARI), paper presented on panel, "Job Performance Measurement," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Multi-Dimensional Performance Measurement. 9

Lawrence M. Hanser and Jane M. Arabian (ARI), and Lauress Wise (AIR), paper presented on symposium, "Issues in Job Performance Measurement," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Measuring Personal Attributes: Temperament, Biodata, and Interests. . . 17

Leaetta M. Hough, Matt K. McGue, John D. Kamp, Janis S. Houston, and Bruce N. Barge (PDRI), paper presented on symposium, "Predicting a Broad Variety of Criteria: Elaborating the Predictor Space," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Computerized Assessment of Perceptual and Psychomotor Abilities. . . . 25

Jeffrey J. McHenry and Jody L. Toquam (PDRI), paper presented on symposium, "Predicting a Broad Variety of Criteria: Elaborating the Predictor Space," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Examination of Environmental Determinants of Army Performance Criteria . 33

Darlene M. Olson (ARI) and Walter C. Borman (PDRI), paper presented on panel, "Organizational Effectiveness," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Mapping Predictors to Criterion Space: Overview 43

Norman G. Peterson (PDRI), paper presented on symposium, "Predicting a Broad Variety of Criteria: Elaborating the Predictor Space", at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Using Microcomputers for Assessment: Practical Problems and Solutions . 51

Rodney L. Rosse and Norman Peterson (PDRI), paper presented on symposium, "Predicting a Broad Variety of Criteria: Elaborating the Predictor Space," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

CONTENTS

Page

The Validity of ASVAB for Predicting Training and SQT Performance. . . . 59

Paul G. Rossmeissl (ARI), Donald H. McLaughlin, Laurence L. Wise, and David A. Brandt (AIR), paper presented on panel, "Entry Standards Based on the ASVAB," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Developing New Attribute Requirements Scales for Military Jobs 67

Elizabeth P. Smith (ARI), paper presented on symposium, "Determining Ability Requirements," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Adding to the ASVAB: Cognitive Paper-and-Pencil Measures. 75

Jody L. Toquam, Marvin D. Dunnette, VyVy A. Corpe, and Janis Houston (PDRI), paper presented on symposium, "Predicting a Broad Variety of Criteria: Elaborating the Predictor Space," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Three Variables That May Influence the Validity of Biodata. 83

Clinton B. Walker (ARI), paper presented on panel, "Utilization and Validation of Biodata," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Leaders' Behavior and the Performance of First Term Soldiers 91

Leonard A. White, Ilene F. Gast, and Michael G. Rumsey (ARI), paper presented on panel, "Dimensions of Leadership," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Vocational Interests as Predictors of Army Performance 99

Hilda Wing (ARI), Bruce N. Barge and Leaetta M. Hough (PDRI), paper presented on symposium, "Elements of a Military Occupational Explorational System," at the Annual Conference of the Military Testing Association, San Diego, October 1985.

Performance Criterion Measurement: What Are the Different Methods Measuring? 107

Walter C. Borman (PDRI), paper presented at the Air Force Conference on Job Performance Measurement, Air Force Human Resources Laboratory, San Antonio, August 1986.

CONTENTS

Page

Standard Setting Procedures: Army Enlistment Standards and Job Performance. 121

Jane M. Arabian and Lawrence M. Hanser (ARI), paper presented on symposium, "Setting Standards in Performance Measurement," at the Annual Convention of the American Psychological Association, Washington, August 1986.

Toward a General Model of Soldier Effectiveness: Focusing on the Common Elements of Performance 165

Walter C. Borman, Elaine D. Pulakos, and Stephen J. Motowidlo (PDRI), paper presented on symposium, "Multiple Criteria and Multiple Jobs: Will One Model Fit All?", at the Annual Convention of the American Psychological Association, Washington, August 1986.

Project A: When the Textbook Goes Operational 187

John P. Campbell (HumRRO), invited address at the Annual Convention of the American Psychological Association, Washington, August 1986.

Comparability of Hands-On and Knowledge Tests Across Nine Military Jobs. 219

Patrick Ford and Charlotte H. Campbell (HumRRO), Daniel B. Felker and Dorothy C. Edwards (AIR), and Michael G. Rumsey (ARI), paper presented on symposium, "Multiple Criteria and Multiple Jobs: Will One Model Fit All?", at the Annual Convention of the American Psychological Association, Washington, August 1986.

The Relation of Leadership and Individual Differences to Job Performance 235

Leaetta M. Hough (PDRI), Ilene F. Gast and Leonard A. White (ARI), and Rodney McCloy (PDRI), paper presented on symposium, "Individual Differences, Work Environment, and Performance in Military Jobs," at the Annual Convention of the American Psychological Association, Washington, August 1986.

Stability and Instability of Individual Differences. 263

Lloyd G. Humphreys (University of Illinois, member of Project A Scientific Advisory Group), paper presented on symposium, "Promising Areas of Psychometric Research," at the Annual Convention of the American Psychological Association, Washington, August 1986, and at the meeting of the Scientific Advisory Group, September 1986.

CONTENTS

Page

Individual Differences and Environmental Determinants of Army Performance Criteria 275

Darlene M. Olson (ARI), Walter C. Borman, and Stephen J. Motowidlo (PDRI), paper presented on symposium, "Individual Differences, Work Environment, and Performance in Military Jobs," at the Annual Convention of the American Psychological Association, Washington, August 1986.

Weighting Performance Constructs in Composite Measures of Job Performance 319

Robert Sadacca, Maria Veronica de Vera, and Ani DiFazio (HumRRO), and Leonard A. White (ARI), paper presented on symposium, "Multiple Criteria and Multiple Jobs: Will One Model Fit All?", at the Annual Convention of the American Psychological Association, Washington, August 1986.

A Path Analytic Model of Job Performance Ratings 381

Leonard A. White (ARI), Walter C. Borman and Leaetta M. Hough (PDRI), and R. Gene Hoffman (HumRRO), paper presented on symposium, "Causal Models of Job Performance," at the Annual Convention of the American Psychological Association, Washington, August 1986.

A Latent Structure Model of Job Performance Factors. 413

Lauress L. Wise (AIR), John P. Campbell (HumRRO), Jeffrey J. McHenry (PDRI), and Lawrence M. Hanser (ARI), paper presented on symposium, "Multiple Criteria and Multiple Jobs: Will One Model Fit All?", at the Annual Convention of the American Psychological Association, Washington, August 1986.

Army Research to Link Standards for Enlistment to On-the-Job Performance 447

Jane Arabian and Michael Rumsey (ARI), and Jeffrey McHenry (AIR), in "Fifth Annual Report to Congress," Army Research Institute Working Paper, September 1986.

IMPROVING THE SELECTION, CLASSIFICATION, AND UTILIZATION OF ARMY ENLISTED PERSONNEL:

ANNUAL REPORT, 1986 FISCAL YEAR SUPPLEMENT TO ARI TECHNICAL REPORT 792

PURPOSE OF THE REPORT

The materials presented in this report were prepared under Project A, the U.S. Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. This Research Note supplements ARI Technical Report _____, the Project Annual Report for the 1986 Fiscal Year. It augments that report by providing copies of a set of technical papers that were prepared during the year reporting on detailed phases of the project research methods and results.

OVERVIEW OF PROJECT A

Project A is a comprehensive long-range research and development program the U.S. Army has undertaken to develop an improved personnel selection and classification system for enlisted personnel. The Army's goal is to increase its effectiveness in matching first-tour enlisted manpower requirements with available personnel resources, through use of new and improved selection/classification tests that will validly predict carefully developed measures of job performance. The project addresses the Army's 675,000-person enlisted personnel system encompassing several hundred military occupations.

The program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research needed to develop the desired system. In 1982 ARI selected a consortium, led by Human Resources Research Organization (HumRRO) and including American Institutes for Research (AIR) and Personnel Decisions Research Institute (PDRI), to undertake the 9-year project. It is utilizing the services of 40 to 50 ARI and consortium researchers working collegially in a variety of professional specialties. The Project A objectives are to:

- o Validate existing selection measures against both existing and project-developed criteria (including both Army-wide job performance measures based on rating scales, and direct hands-on measures of MOS-specific task performance).

- o Develop and validate new selection and classification measures.

- o Validate intermediate criteria such as training performance, as predictors of later criteria, such as job performance, so that better informed decisions on reassignment and promotion can be made throughout a soldier's career.

- o Determine the relative utility to the Army of different performance levels across MOS.

- Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making decisions.

The research design incorporates three main stages of data collection and analysis in an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria). In the first iteration, file data from fiscal years (FY) 1981/1982 were evaluated to explore relationships between scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB), and their later performance in training and their scores on first-tour Skill Qualification Tests (SQT).

For the ensuing research, 19 Military Occupational Specialties (MOS) were selected as a representative sample of the Army's 250+ entry-level MOS. The selection was based on an initial clustering of MOS derived from rated similarities of job content. These MOS account for about 45 percent of Army accessions and provide sample sizes large enough so that race and sex fairness can be empirically evaluated in most MOS.

In the second iteration, a Concurrent Validation design was executed with FY83/84 accessions. A "Preliminary Battery" of perceptual, spatial, temperament, interest, and biodata predictor measures was developed and tested with several thousand soldiers as they entered four MOS. The data from this sample were then used to refine the measures, with further exploration of content and format. The revised set of measures was field tested to assess reliabilities, "fakability," practice effects, and other factors. The resulting predictor battery, the "Trial Battery," was administered together with a comprehensive set of job performance indexes based on job knowledge tests, hands-on job samples, and performance rating measures, in the Concurrent Validation during the summer and fall of 1985.

On the basis of testing experience, the "Trial Battery" was revised as the "Experimental Predictor Battery," which in turn is being administered in the Longitudinal Validation stage (third iteration), beginning in the late summer of 1986. All measures are being administered in a true predictive validity design. About 50,000 soldiers across 21 MOS will be included in the FY86-87 administration and subsequent first-tour measurement. About 3,500 of these soldiers are expected to be available for second-tour performance measurement in FY91. Three MOS have been added to the original 19 (19K, 29E, 96B), and one of the original MOS was dropped (76W).

Activities and progress during the first three years of Project A were described in annual reports as follows: FY83 - ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37; FY84 - ARI Research Report 1393 and related reports, ARI Technical Report 660 and ARI Research Note 85-14; FY85 - ARI Technical Report _____ (in preparation) and ARI Research Note ___ (in preparation). These reports list other publications on specific activities.

Other publications on specific activities during those years are listed in those annual reports. The annual report on project-wide activities during FY86 is presented in ARI Technical Report ______. The technical papers reproduced in this Research Note serve as additional documentation for various FY86 activities.

UTILITY ESTIMATION IN FIVE ENLISTED OCCUPATIONS

Newell K. Eaton
Hilda Wing
Alan Lau

U.S. Army Research Institute

Presented on symposium,
"Job Performance Measurement"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Utility Estimation in Five Enlisted Occupations

Newell K. Eaton, Hilda Wing and Alan Lau

U.S. Army Research Institute for the Behavioral and Social Sciences[1]

In most organizations the decision to develop and implement selection and/or classification tests rests on the assumption that their costs will be outweighed by their benefits in terms of increased employee performance and tenure. The initial costs of testing programs have been increasing due to more stringent requirements for documentation of validities, test administration using computers, and the potential for legal challenges to test fairness. With the increasing costs of starting and maintaining testing programs, more attention is being paid to assessing their benefits. The purpose of this paper is to expand on methods used by several researchers in this area (Eaton, Wing, & Mitchell, 1985; Hunter & Schmidt, 1982).

Brogden (1949) and Cronbach and Gleser (1965) provided the first systematic descriptions of the utility of testing programs indexed in dollars. They linked performance levels to the dollar values estimated for those performance levels. Their formula for the gain in productivity, or utility (U$), obtained by using valid selection procedures includes (a) Ns, the number of individuals selected; (b) SD$, the standard deviation of performance, scaled in a utility metric such as dollars; and (c) the average performance expected on the criterion by the selected group as estimated from a valid predictor, given by Rxy Zx:

$$U\$ = Ns\ SD\$\ Rxy\ Zx$$

The formula was subsequently modified to account for testing costs. A more complete description of such formulations can be found in Cascio (1982), Cronbach and Gleser (1965), and Hunter and Schmidt (1982).

While the values of most of the variables on the right hand side of the Brogden-Cronbach-Gleser formulas are known, the estimation of SD$, the standard deviation of performance scaled in dollars, is problematic. One 'SD$ Estimation Technique' is based on estimates of the dollar value to the organization of performance at the 50th percentile level, the 85th percentile level (one standard deviation above the mean), and, sometimes, the 15th percentile level (one standard deviation below the mean). The dollar difference between the 15% and 50% estimates, and the 50% and 85% estimates, provides an estimate of SD$ (Cascio & Silbey, 1979; Hunter & Schmidt, 1982; and Schmidt, Hunter, McKenzie, & Muldrow, 1979).

----

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

A second method is the 'Superior Equivalents Technique' proposed by Eaton et al.(1985). It is somewhat like the SD$ Estimation Technique. Instead of using estimates of the dollar value of 85th percentile performance, however, the technique uses estimates of the number (N85) of superior (85th percentile) performers who would be needed to produce the output of a fixed number (N50) of average (50th percentile) performers. This estimate, combined with an estimate of the dollar value (V50) of average performance, provides an estimate of SD$:

$$SD\$ = V50\ [(N50/N85) - 1].$$

Eaton et al. speculated that this method would be more appropriate in situations where the nature of the work is such that managers are more accustomed to considering the relative productivity of employees or crews than the relative costs of producing given levels of output.

A third estimation strategy has been proposed by Hunter and Schmidt (1982). In reviewing the results of a variety of studies, they note that SD$ typically falls between 40% and 70% of annual salary. This might be termed the 'Salary Percentage Technique.'

In their recent paper, Eaton et al. showed that the Superior Equivalents Technique provided more stable estimates of U.S. Army tank commanders' SD$ than did the SD$ Estimation Technique. They also noted that both these techniques provided substantially larger estimates of SD$ than did the Salary Percentage Technique. The purpose of this paper was to compare the results of the Superior Equivalents Technique with the SD$ Estimation Technique across five different U.S. Army enlisted military occupational specialties (MOS). This was intended to assess both the variability of SD$ values across the five MOS as well as the results with the two techniques. The paper was also intended to determine whether a 'short hand' estimation procedure could be developed for military occupations, such as the Salary Percentage Technique. Last, because the research was conducted with supervisors who were both noncommissioned officers (NCOs) and commissioned officers, it was possible to assess the impact of level of management on SD$ estimates.

METHOD

Instrument

A questionnaire based on earlier research (Bobko et al. 1983; Burke & Frederick, 1984; Eaton et al. 1985; Schmidt et al. 1979) was developed to measure the comparative worth to the Army of first-term soldiers operating at different performance levels. Separate forms were administered to supervisors in each of the five MOS studied. The first method asked supervisors to think about how much ten average soldiers (50th percentile) contributed to the Army. Supervisors then estimated how many superior (85th percentile) soldiers would be needed to do the same amount of work. The second method asked supervisors to first consider the worth of average and superior first tour soldiers to the Army. They were then asked to estimate how much an average (50th percentile) first-term soldier and a superior (85th percentile) soldier are worth by considering such factors as salary, output, responsibility, and equipment. Dollar estimates of the yearly value to the Army of average and superior soldiers were then requested.

Subjects

Supervisory estimates were obtained from 270 NCOs and officers in five MOS across three different posts. The five MOS were infantrymen (11B), armor crewmen (19B), light wheel vehicle/power mechanics (63B), medical specialists (91B), and radio teletype operators (05C). Of the 270 supervisors, 226 (83 percent) were NCO and 29 (11 percent) were officers. The remainder did not provide rank information. Four supervisors (one percent) did not respond to the methods of estimating utility and their responses are not included in the analyses. Of the remaining 266, 13 did not provide useable estimates for the first method and (a different) eight did not provide useable estimates for the second method.

Other Data

To obtain the value of average performance for the Superior Equivalents Technique, as well as the data required for the Salary Percentage Technique, we used published pay and allowance tables. In 1985 the base pay for Army enlisted personnel with two years of service ranged from $9,000 to $10,000. Non-taxable allowances for such items as housing, post exchange, vacation, and travel benefits could amount to more than $6,000 for the typical married soldier with dependents. Our estimate of an equivalent civilian salary would be about $16,000 per year. This is consistent with Henderson's (1985) estimates for the compensation of a Private First Class living off post with dependents.

RESULTS

The results from the Superior Equivalents Techniques indicated that, across MOS, 5.20 superior first-tour soldiers performed as well as 10 average soldiers. Using $16,000 as the value of average performance (V50), 5.20 as the number of superior equivalents (N85), and 10 as the number of average soldiers (N50), the Superior Equivalents Technique yielded a SD$ estimate of $14,769. Of the 253 supervisors responding, 7% indicated 1 or 2 superior first-tour enlisted soldiers were equivalent to 10 average soldiers, 23% indicated 3 or 4, 51% indicated 5 or 6, 17% indicated 7 or 8, and 3% responded with 9 or 10. There was only a modest difference between estimates for the five MOS: the number of superior equivalents ranged from 4.90 to 5.58 with SD$ estimates from $12,881 to $16,720. The results by MOS are shown in Table 1. Full ANOVA results were computed, including as factors MOS and RANK of the supervisor providing the estimates. The differences by MOS did not reach statistical significance, nor did RANK, nor the MOS x RANK interaction.

The results from the SD$ Estimation Technique indicated that, across MOS, average soldiers were worth about $16,725 per year while superior soldiers were worth about $25,969. This yields an SD$ estimation of $9,244. Of the 258 supervisors responding, 11% provided SD$

Table 1: Estimated Number of Superior First Tour Soldiers Equaling 10 Average Soldiers and Computed SD$ by MOS

MOS	N	Number Superior	SD$
11B	48	5.54	$12,881
19E	60	5.40	$13,630
91B	36	4.89	$16,720
63B	67	5.15	$15,068
05C	42	4.90	$16,653
Totals	253	5.20	$14,769

estimates of less than $2,000, 14% between $2,001 and $4,000, 19% between $4,001 and $6,000, 12% between $6,001 and $8,000, 16% between $8,001 and $10,000, 15% between $10,001 and $16,000, and 12% over $16,000. These appear to be more variable than Superior Equivalents estimates. Larger, between MOS differences also were found with the SD$ estimation technique, ranging from about $6,254 to $11,150. The average values assigned average and superior soldiers, as well as SD$ estimates for the five MOS, are shown in Table 2.

Table 2: Dollar Estimates of Value to the Army of Average and Superior First Tour Soldiers by MOS

MOS	N	Average	Superior	SD$
11B	53	$19,226	$29,000	$ 9,774
19E	63	13,736	20,190	6,254
91B	38	18,000	27,132	9,132
63B	64	15,719	26,344	10,625
05C	40	18,200	29,350	11,150
Totals	258	$16,725	$25,969	$ 9,244

Full ANOVA results were computed on the SD$ estimates following the procedures outlined for the Superior Equivalents estimates. With the SD$ estimates, however, the effect of MOS was significant (F = 4.23, df = 4,225, p < .01). Duncan's multiple range tests indicated the SD$ estimates for first tour armor crewmen (19E) were lower than those for medics (91B) mechanics (63B) and radio telephone operators (05C). Neither the RANK nor MOS x RANK effects were significant.

Last, SD$ values obtained using both the Superior Equivalents and SD$ Estimation Techniques were compared to the estimated civilian equivalent salary and to base pay. Using $16,000 as the best estimate of estimated civilian equivalent salary and $9,500 as base pay, estimates of SD$ would be 58-92% of estimated civilian equivalent salary based on superior equivalents and SD$ estimates, respectively. Using only base pay as a salary basis, SD$ would be estimated at 97%-156%. The value of 125% of base pay may be chosen as an estimate of SD$. Assuming a value of $10,000 per year as base pay (for simplicity, rather than the $9,500 figure used in previous analyses), then SD$ = $12,500 and U$ can be estimated (Cascio, 1982, pp 220-226). Table 3 displays the estimated U$, per first tour soldier selected, as a function of the validity of the test and the proportion of applicants selected.

Table 3: Estimated U$ Per Selection as a Function of Test Validity and Proportion of Applicants Selected

		Test Validity					
		.1	.2	.3	.4	.5	.6
Proportion	.2	$1,750	$3,500	$5,250	$7,000	$8,750	$10,500
of	.4	1,200	2,400	3,600	4,800	6,000	7,200
Applicants	.6	813	1,625	2,438	3,250	4,063	4,875
Selected	.8	413	825	1,238	1,650	2,063	2,475

If 100 soldiers were selected from among 125 applicants, using a test with a validity of .3, the estimated utility would be 100 x $1,238 = $123,800 per year.

DISCUSSION

The first purpose of this research was to assess the SD$ of performance in five Army enlisted military occupational specialties using two methods. For both methods there were numerical differences in SD$ across the MOS, and they were ordered logically. The lowest SD$ values were obtained for team/crew occupations - infantryman and tank crewman - while the highest SD$ values were obtained for those who perform many duties as individuals - medics, mechanics, and radio/telephone operators. However, between MOS differences were statistically significant for SD$ values obtained for only one method, the SD$ Estimation Technique, and these differences were not clear cut.

The Superior Equivalents Technique, designed for use in military settings, did not provide reliable between-MOS differences. It did, however, yield estimates with considerably smaller levels of between-subjects dispersion. This is consistent with the results of the earlier Eaton et al. research. On balance, it would seem that both techniques provide SD$ estimates which yield a useful range in which the 'real' SD$ probably falls. But neither is sufficiently precise at this time to provide between-MOS comparisons in which one can be confident.

Obtaining a ball-park estimate of SD$ may well be sufficient for most purposes. Seldom does one face a decision where the utilization of a selection or classification test rests on cost tradeoffs of plus or minus 10% or 20% of testing and start up costs. Rather, such programs are more typically initiated only if the potential payoff is several times the costs. As a consequence, estimating a reasonable range of SD$ values can be quite useful.

Fortunately, this and prior research (Eaton et al. 1985) show that such an estimate may be obtained using a variant of the Hunter & Schmidt (1982) Salary Percentage Technique. In the Eaton et al. work, SD$ was 89% of estimated civilian equivalent salary, and 178% of base pay. For the two methods compared in this research, results ranged from 58%-92% of civilian equivalent salary, and 97%-156% of base pay. Given this consistency it would seem that a rough estimate of SD$ for first-tour enlisted personnel is about 125% of base pay.

Such an estimate is likely to be quite conservative. Eaton et al. found SD$ values obtained with the two methods used in this research to be about half those obtained with yet a fourth method, the System Effectiveness Technique, designed to incorporate equipment, maintenance, and other support costs. Burke and Frederick (1984) and Schmidt et al. (1982) also obtained results suggesting the conservative nature of SD$ values obtained with the SD$ Estimation Technique. The use of such a rough estimate may well make a useful contribution to front end analyses designed to assess the potential utility of initiating research on, or implementation of, a selection and classification testing program. Table 3 provides figures which make such estimates relatively simple.

REFERENCES

Bobko, P., Karren R., & Parkington, J.J. (1983). The estimation of standard deviations in utility analyses: An empirical test. Journal of Applied Psychology, 68, 170-176.

Brogden, H.E. (1949). When testing pays off. Personnel Psychology. 2, 171-183.

Burke, M.J., & Frederick, J.T. (1984). Two modified procedures for estimating standard deviations in utility analyses. Journal of Applied Psychology, 69, 482-489.

Cascio, W.F. (1982). Costing human resources: The financial impact of behavior in organizations. Boston: Kent Publishing Co.

Cascio, W.F., & Silbey, V. (1979). Utility of the assessment center as a selection device. Journal of Applied Psychology, 64, 107-118.

Cronbach, L.J., & Gleser, G.C. (1965). Psychological tests and personnel decisions (2nd ed.), Urbana: University of Illinois Press.

Eaton, N.K., Wing, H., & Mitchell K.J. (1985). Alternate methods of estimating the dollar value of performance. Personnel Psychology. 38, 27-40.

Henderson, W.D. (1985). Cohesion: The human element in combat. Washington, D.C., National Defense University Press.

Hunter, J.E., & Schmidt, F.L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In Fleishman E.A., Dunnette M.D. (Eds.), Human performance and productivity: Volume I. Human capability assessment. Hillsdale, NJ,: Erlbaum.

Schmidt, F.L., Hunter, J.E., McKenzie, R., & Muldrow, T. (1979). The impact of valid selection procedures on workforce productivity. Journal of Applied Psychology, 64, 609-626.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work force productivity. Personnel Psychology. 35, 333-347.

MULTI-DIMENSIONAL PERFORMANCE MEASUREMENT

Lawrence M. Hanser and Jane M. Arabian
U.S. Army Research Institute

Lauress Wise
American Institutes for Research

Presented on symposium,
"Issues in Job Performance Measurement"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Multi-dimensional Performance Measurement

Lawrence M. Hanser and Jane M. Arabian
U.S. Army Research Institute[1]

Lauress Wise
American Institutes for Research

Introduction and Background

This paper is based on data collected for the large Army personnel research project titled "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Project A" (Eaton, Hanser, & Shields, 1985). This project was conceptualized and planned during the 1980 to 1981 time period, and a contract was signed with the prime contractor, Human Resources Research Organization (HumRRO), in 1982. It is being conducted jointly by scientists from the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), HumRRO, the American Institutes for Research (AIR), and Personnel Decisions Research Institute (PDRI).

Early in the planning for Project A, it was recognized that a large proportion of the research would have to be devoted to criterion development. Plans called for the development of several different measures of performance: (a) tests of hands-on performance, (b) paper and pencil tests of job knowledge, and (c) ratings of typical performance. Each of these broad categories of criteria were further subdivided. Hands-on tests included tasks which were specific to each Military Occupational Specialty (MOS) as well as tasks common to all MOS. Two kinds of paper and pencil tests were constructed: (a) to emphasize the content of formal school training, and (b) to emphasize MOS-specific task performance. Rating forms were constructed both for MOS-specific task performance as well as for non MOS-specific Army-wide performance that we have labelled broadly as "soldiering."

The initial impetus for developing such a comprehensive set of criterion measures was largely a function of our underlying theory of performance measurement. This underlying theory states rather simply that performance in a job is multi-dimensional, and that it is not possible to capture that multi-dimensionality using only one measurement method. A method of measurement may be intrinsic to some tasks. For example, having the requisite knowledge of how to take a person's blood pressure may not be the same as actually being able to perform the task accurately, yet both are important. An individual may score high on a paper and pencil test on this task, but might not score as high on a hands-on test of this task. In order to be successful in performing this task on the job it requires: (a) the knowledge of how to do the task, (b) the physical skills to perform the task, and (c) the motivation to do it. Or to put it in another well known way: performance = f(ability x motivation).

[1] The views expressed in this paper are those of the author and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army

Because of the complexity of the criterion space being measured in this project, it is extremely important that it be fully understood prior to choosing a final set of predictors and recommending changes to the Army's selection and classification procedures. Several recent papers by project scientists have begun to address the issues associated with criterion development (c.f., Borman, White, Gast, & Pulakos, 1985; Campbell & Harris, 1985; Rumsey, Osborn, & Ford, 1985). Borman et al. constructed and tested a path model of supervisory and peer ratings to examine how each are related to other measures of performance. They found that both job knowledge and hands-on task proficiency are related to ratings, with the dominant path between ratings and hands-on proficiency. They conclude, however, that "... for the most part different methods of measuring job performance yield quite different results." Campbell and Harris describe the results of attempting to interpret criteria using a group of "concerned psychologists." They also present a "working model of job performance for the domain of skilled jobs." In examining the correlation matrices of hands-on and job knowledge tests and rating scales, they state "... the methods correlate more highly within themselves than they do across measures." Rumsey et al. examine the relationships between job knowledge tests and hands-on tests of job proficiency. In each of these papers, a central theme is the multi-dimensionality of performance and the importance of using different measurement methods to capture performance adequately.

The intent of this paper is to further explore the criterion space measured in Project A. Previous research has focused on aggregate measures of performance such as total scores on hands-on or paper and pencil tests or average ratings across several dimensions. In this paper we focus on task level measures in order to begin to understand better the relationships between kinds of tasks and methods of measuring performance on them. Through this we hope to gain a better understanding of the method variance associated with measures of task performance.

Method

Subjects

Data reported in this paper were collected in 1984 as part of field tests of the criterion measures developed by Project A scientists. Participants included first tour soldiers in two Army MOS: (a) 178 Infantrymen (MOS 11B) and (b) 167 Medical Specialists (MOS 91A). A complete description of the data collection methods can be found in Campbell and Harris (1985).

Variables

Percent correct steps per task and average supervisory rating per task provided the major variables used in these analyses. Percent correct scores were obtained on both hands-on and written tests. For each MOS reported here, approximately 15 tasks were scored using all three measurement methods: (a) hands-on performance, (b) multiple choice paper and pencil test, and (c) average supervisory rating of task performance. Approximately 15 additional tasks per MOS were tested in the paper and pencil test, and these were also included in the analyses. In addition, total score on a paper and pencil test focusing on training course content, average supervisory rating on overall performance, and Armed Services Vocational Aptitude Battery (ASVAB) subtest standard scores were included. This resulted in a total of approximately 71 variables per MOS to be included in these analyses. Although these are a relatively small number of subjects given the number of variables, the limits of analysis are a function of the number of factors extracted. These sample sizes will support the extraction of a maximum of five to seven factors per MOS.

Analyses

Though some "feel anxious in the presence of too many partial or semi-partial correlations" (Campbell & Harris, 1985), we decided to explore these data using factor analysis. Our specific plans were as follows: (a) extract a set of oblique factors for each MOS, (b) examine the inter-factor correlation matrices, and (c) examine the patterns of loadings within and across MOS. We used a principal axis solution with an iterative solution for the communalities and a Promax rotation. We decided on the number of factors to extract based on an inspection of the scree and interpretability of various solutions. In order to conserve space, descriptive statistics and reliabilities are not reported here. They are, however, available elsewhere (Borman et al., 1985; Campbell & Harris, 1985; Rumsey et al., 1985).

Results and Discussion

The data on the Medical Specialists yielded a five factor solution. Table 1 shows the oblique solution. Variables reported in the table are limited to the three highest loading on any factor, any variable with an absolute loading of greater than .30 on a cross-method factor, and any variable with loadings greater than .30 on two or more factors.

Table 1. Rotated Factor Pattern (STD REG COEFS)

I	II	III	IV	V	
80	.	.	.	.	Rating:Splint Suspected Fracture <Supv>
77	.	.	.	.	Rating:Put on Field/Pres Dressing <Supv>
75	.	.	.	.	Rating:Perform CPR <Supv>
58	.	.	-35	.	Rating:Measure/Record Respir. <Supv>
53	.	30	.	.	Rating:Measure/Record Pulse <Supv>
.	57	.	.	.	P&P:D9-Replace Filters in M17 Mask
.	51	.	.	.	P&P:I4-Measure/Record Respirations
.	47	.	.	.	P&P:I9-Estab/Maintain a sterile fld
.	43	.	.	.	HO: A4-Put on Field/Pres Dressing
.	34	.	.	.	HO: A9-Init a Field Med Card
.	.	68	.	.	ASVAB SUBTEST SCR-Arithmetic Reasoning
.	.	57	.	.	ASVAB SUBTEST SCR-Math Knowledge
.	.	52	.	.	ASVAB SUBTEST SCR-Coding Speed
.	.	49	.	.	P&P: I6-Assemble Needle & Syringe
.	.	49	.	.	P&P: K2-Draft/Fire TPR Charts
.	.	42	.	.	P&P: A6-Open Airway
.	.	40	.	.	P&P: I7-Change a Sterile Dressing
.	.	41	32	.	School: All Items
.	.	.	76	.	ASVAB SUBTEST SCR-Auto/Shop
.	.	.	71	.	ASVAB SUBTEST SCR-Electronics Information
.	.	.	59	.	ASVAB SUBTEST SCR-Mechanical Comprehension
.	.	.	37	.	P&P: G3-Vehicle Recognition
.	.	.	.	68	HO: I3-Measure/Record Pulse
.	.	.	.	51	HO: I9-Est/Maintain Sterile Field

.	.	.	.	47	HO: I4-Measure/Record Respir.
.	.	.	33	35	HO: AB-Splint Suspected Fracture

As expected, there are strong method factors, with little overlap of wariables across method factors. Note, however, that two ratings overlap with the ASVAB factors, and one of the hands-on tasks overlaps with an ASVAB factor. Two hands-on tasks have loadings greater than .30 on Factor II, the paper and pencil job knowledge test factor. Several of the job knowledge test tasks load on the two ASVAB factors. Also, ASVAB splits into two factors, a math/speed factor and a technical factor. Table 2 provides the factor correlations.

Table 2. Inter-Factor correlations

	I	II	III	IV	V
I	100	1	7	-11	17
II	1	100	15	27	-2
III	7	15	100	-6	19
IV	-11	27	-6	100	-8
V	17	-2	19	-8	100

Not surprisingly, the paper and pencil job knowledge test factor, Factor II, and an ASVAB factor, Factor IV, have the highest correlation. Note, however, that none of the ASVAB subtests have loadings of .30 or higher on Factor II, and that it is the ASVAB technical factor which correlates highest with the job knowledge paper and pencil test factor. The ASVAB Verbal subtest did not meet the criteria for inclusion in this table. These results would seem to indicate that correlations between ASVAB and paper and pencil job knowledge measures are not simply the result of shared method variance.

The next highest inter-factor correlations are between the hands-on factor, Factor V, and the ASVAB math/speed and rating factors, Factors III and I respectively. While the hands-on factor is a relatively pure method factor, its correlations with the other factors strengthen the conclusions of Borman et al. Each method appears to measure a different but related piece of job performance.

Table 3 contains the oblique promax factor pattern for Infantrymen. Seven factors were extracted. The choice of variables to report was based on the same rules as for the previous table of loadings.

Table 3. Rotated Factor Pattern (STD REG COEFS)

I	II	III	IV	V	VI	VII	
64	.	.	.	.	.	.	P&P: E5-Oper as Station in Radio Net
64	.	.	.	.	.	.	School: All Items
61	.	.	.	.	.	.	P&P: B4-Perform OP Maint. on M16A1
59	.	.	.	.	.	.	P&P: H1-Perform Tracked Vehicle Maint
56	.	.	-39	.	.	.	P&P: E1-Collect/Report Info
.	66	.	.	.	.	.	Rating: Install/Fire/Recover M18A1 <Supv>
.	65	.	.	.	.	.	Rating: Load/Clear M60 <Supv>
.	59	.	.	.	.	.	Rating: Prepare Range Card for M60 <Supv>
.	54	.	37	.	.	.	Rating: Mean non MOS-Specific<Supv>
.	50	.	33	.	.	.	Rating: Navigate on Ground <Supv>
.	38	.	39	.	.	.	Rating: Set Headspace/Timing on .50 <Supv

.	31	.	.	.	.	.	P&P: G8-Estimate Range
.	.	76	.	.	.	.	ASVAB SUBTEST SCR-Auto/Shop
.	.	74	.	.	.	.	ASVAB SUBTEST SCR-Mechanical Comprehension
.	.	73	.	.	.	.	ASVAB SUBTEST SCR-General Science
.	.	73	.	.	.	.	ASVAB SUBTEST SCR-Verbal
.	.	.	79	.	.	.	Rating: Op as Station in Radio Net <Supv>
.	.	.	76	.	.	.	Rating: Op Radio Set AN/PRC-77 <Supv>
.	.	.	44	.	.	.	HO: E5-Op as Station in Radio Net
.	.	.	31	.	.	.	HO: BC-Engage Targets w LAW
.	.	.	.	68	.	.	HO: C6-Call/Adjust Indirect Fire
.	.	.	.	67	.	.	HO: G8-Estimate Range
.	.	.	.	55	.	.	HO: B4-Perform Op Maint on M16A1
.	39	.	.	37	.	.	Rating: Call/Adjust Indirect Fire <Supv>
30	.	.	.	32	.	.	P&P: B9-Engage w Hand Grenades
32	.	.	.	.	.	.	HO: BB-Prepare Range Card for M60
.	.	.	.	.	58	.	HO: J1-Movement in Urban Terrain
.	.	.	.	.	56	.	HO: BA-Prepare Dragon for Firing
.	.	.	.	36	50	.	HO: B9-Engage Targets w Grenades
.	.	.	.	.	47	.	HO: I1-Install/Fire/Recover M18A1
.	.	.	.	.	35	.	P&P: BA-Prepare Dragon for Firing
.	.	.	.	.	.	71	ASVAB SUBTEST SCR-Numerical Operations
.	.	.	.	.	.	59	ASVAB SUBTEST SCR-Coding Speed
.	.	40	.	.	.	54	ASVAB SUBTEST SCR-Math Knowledge
.	.	41	.	.	.	53	ASVAB SUBTEST SCR-Arithmetic Reasoning

While similar method factors emerge, the factor space for infantrymen is slightly more complex. The ASVAB Factors III and VII are quite clean, though Factor III and the paper and pencil job knowledge test Factor I are relatively oblique (Table 4.). These factors are substantially more correlated than are the two ASVAB factors with each other. Note also that the ASVAB math/speed Factor VII has a lower correlation with the paper and pencil job knowledge test Factor I, than the more technical ASVAB Factor III. If there is a simple "written test" factor, it failed to emerge in either of these solutions.

Perhaps most interesting are Factors IV and V. Each of these factors has a mixture of variable loadings representing different measurement methods. On Factor IV the supervisory rating and hands-on test for operating as a radio station in a net both load substantially. On Factor V the supervisory rating and hands-on test for call/adjust indirect fire both load substantially, and the paper and pencil and hands-on tests for engage targets with grenades also both load substantially.

Table 4 gives the correlations among the factors for Infantrymen. This solution is considerably more oblique than the solution for Medical Specialists.

Table 4. Inter-Factor correlations

	I	II	III	IV	V	VI	VII
I	100	30	53	36	18	25	24
II	30	100	13	40	18	19	-3
III	53	13	100	6	13	-1	29
IV	36	40	6	100	21	34	-5
V	18	18	13	21	100	-2	1

VI	25	19	-1	34	-2	100	0
VII	24	-3	29	-5	1	0	100

The highest correlation is between Factor I, the paper and pencil job knowledge test factor, and Factor III, the ASVAB technical factor. This result is similar to that noted previously. The two primarily supervisory rating factors, II and IV, are quite highly correlated with the paper and pencil test of job knowledge factor. In fact, Factor IV correlates almost as highly with Factor I (r=.36) as it does with the other rating factor, Factor II (r=.40). The two hands-on test factors, V and VI, are uncorrelated with each other. Factor VI has respectable correlations with both the paper and pencil job knowledge test factor, Factor I, and the rating factor, Factor IV.

Conclusions

Our tendency as psychologists is to abhor method variance as something to be avoided. This should not necessarily be the case in the realm of job performance measurement. Performance of a task requires first the ability and motivation to learn the task, and second the skill, ability, and motivation to perform it. Different methods of measuring performance, hands-on tests, written tests, and ratings, capture slightly different aspects of performance. Some of these relationships are apparent from the data presented above.

What remains for us is to understand which kinds of tasks are most appropriately measured by which methods. The research reported here, while open to several interpretations, presents a method and several examples of a way to do this. Clearly, more research needs to be conducted into the content of the tasks themselves and their relationships to method factors across several more occupations than are included here.

References

Borman, W. C., White, L. A., Gast, I. F., & Pulakos, E. D. (1985, August). Performance Ratings as Criteria: What is being Measured. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Campbell, J. P., & Harris, J. H. (1985, August). Criterion Reduction and Combination via a Participative Decision Making Panel. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Eaton, N. K., Hanser, L. M., & Shields, J. L. (in press). Validating Selection Tests Against Job Performance. In J. Zeidner (ED.), Human Productivity Enhancement. New York: Praeger.

Rumsey, M. G., Osborn, W. C., & Ford, P. (1985, August). Comparing Work Sample and Job Knowledge Measures. Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

MEASURING PERSONAL ATTRIBUTES:
TEMPERAMENT, BIODATA, AND INTERESTS

Leaetta M. Hough, Matt K. McGue, John D. Kamp,
Janis S. Houston, and Bruce N. Barge

Personnel Decisions Research Institute

Presented on symposium,
"Predicting a Broad Variety of Criteria:
Elaborating the Predictor Space"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Measuring Personal Attributes: Temperament, Biodata, and Interests

Leaetta M. Hough, Matt K. McGue, John D. Kamp,
Janis S. Houston, and Bruce N. Barge

Personnel Decisions Research Institute

Overview. I'm going to describe the development and evaluation of temperament, biographical, and interest measures - what we call non-cognitive measures - included in the Project A predictor battery. Non-cognitive measures were included in the predictor battery because of their potential for predicting important on-the-job criteria, criteria such as Effort, Initiative, Following Regulations and Orders, Adjustment, Leadership, and Self-Control.

The information I will present today suggests: 1) that non-cognitive predictors are likely to predict such criteria; in fact, more likely to predict such criteria than are other types of predictors; 2) that non-cognitive measures contribute unique variance to the predictor battery and are, therefore, likely to contribute incremental validity; 3) that the non-cognitive measures we developed have good psychometric characteristics, they are internally consistent and show high test-retest reliability; and 4) that faking on personality inventories is not the problem it is often assumed to be. Our overall strategy was: to review the literature on temperament, biodata, and interest to identify constructs that were likely to be criterion valid; to obtain expert judgments about expected true validity of those constructs; to develop measures of those constructs; to remove or revise sensitive or objectionable items; and to evaluate and revise measures based on their internal consistency, overlap with other predictors, and their stability across time and different motivational conditions.

Literature Review Results. Our review and summary of the literature indicated that the validity of interest measures for important Army criteria were in the high .20s. The validities of biographical inventories for such criteria were in the .20s and .30s. These results were not too different from previous literature reviews. Our conclusions for the personality literature, however, differ from some of the other reviews, and I'd like to describe these results more thoroughly.

The criterion-related validities reported in the literature for temperament constructs are shown in Table 1. As you can see, the adjustment criterion, which includes such things as unfavorable discharge and drug abuse, is predicted very well by temperament measures. The predictor constructs Achievement and Locus of Control also predict Educational, Training, and Job Proficiency criteria. These results differ from those reported by Guion and Gottier in their 1965 Personnel Psychology article. Our results are, however, similar to those reported by Ghiselli in his 1973 Personnel Psychology article. We believe the results are explained by the approach we used.

Our approach was to develop a predictor taxonomy and to classify temperament scales into the taxon or construct with which they were most similar. We accomplished this classification by searching the literature

Table 1

Summary[a] of Criterion-Related Validities of Temperament Constructs

			Type of Criterion		
Temperament			Job	Job	
Construct	Educational	Training	Proficiency	Involvement	Adjustment
Potency (Surgency)	.06 (42)[b]	.13 (36)	.07 (65)	.04 (13)	-.17 (31)
Adjustment	.14 (43)	.19 (28)	.11 (65)	.17 (16)	$\boxed{-.33}$ (52)
Agreeableness (Likeability)	.03 (9)	.08 (5)	.03 (22)	-.02 (5)	-.03 (5)
Dependability	.13 (24)	.12 (20)	.11 (49)	.14 (15)	$\boxed{-.43}$ (40)
Intellectance (Culture)	.17 (6)	.19 (5)	.01 (16)	-.09 (9)	.18 (3)
Affiliation	-.03 (5)	--- ---	-.02 (6)	.09 (4)	-.07 (4)
Achievement	$\boxed{.30}$ (8)	$\boxed{.33}$ (4)	$\boxed{.24}$ (4)	--- ---	$\boxed{-.33}$ (5)
Masculinity	-.16 (8)	.09 (3)	.10 (10)	.03 (4)	-.13 (11)
Locus of Control	$\boxed{.32}$ (1)	$\boxed{.29}$ (2)	$\boxed{.25}$ (7)	--- ---	--- ---
Unclassified Military Scales	--- ---	.18 (8)	.18 (25)	--- ---	$\boxed{-.22}$ (20)

[a] Medians are reported as the summary index.
[b] The number in parentheses is the number of correlations on which the median is based.
NOTE: Median correlations greater than .20 are indicated by a box.

for reported correlations between temperament scales and then using these correlations to categorize the temperament scales into the five factors identified by Tupes and Christal (1961) in their peer rating research. We then added four constructs to the taxonomy to increase the homogeneity of the constructs. We also used a taxonomic system for the criteria. These consisted of Educational, Training, Job Proficiency, and Adjustment criteria.

We then summarized the criterion-related validities reported in the literature according to our predictor and criterion taxonomies. Guion and Gottier did not summarize the literature according to constructs; Ghiselli, however, reported results only for studies for which he felt the predictor was conceptually appropriate for the criterion. Our literature review, which summarized the reported validities according to a data- ised classification of scales into constructs, supports Ghiselli's results and conclusions. We believe the construct approach highlighted the predictor-criterion relationships by reducing the "noise," if you will, and that the Guion and Gottier approach masked such relationships.

Expert Judgments of True Validity. Using the construct approach, we identified the temperament constructs that were likely to yield good criterion-related validities. We then asked experts to estimate the expected true criterion-related validities of predictor constructs for important Army criteria. These estimated validities also indicated that the non-cognitive predictors were likely to predict Army criteria - criteria such as Initiative/Effort, Following Regulations and Orders, Leading and Supporting, Self-Control, and others in the .20s, .30s, and even .40s. I might add that the cognitive and psychomotor measures were not expected to predict these criteria nearly as well.

Development of Construct Measures. Using the results of the literature review and expert judgments, we identified "good bets" for predicting important Army criteria. We developed scales to measure these constructs.

We wrote temperament and biodata items for the ABLE, which stands for Assessment of Background and Life Experiences, and we wrote interest and biodata items for the AVOICE, which stands for Army Vocational Interest Career Examination. We also developed four "response validity scales" which we called Social Desirability, Poor Impression, Self-Knowledge, and Non-Random Responses and included the items in these four response validity scales in the ABLE.

We next examined the ABLE and AVOICE items for sensitivity, or the extent to which people might object to the content of the questions. The Army and their scientific advisors also reviewed the items for sensitive content. We revised or removed the objectionable items and administered the ABLE and AVOICE to soldiers at Ft. Lewis, Ft. Campbell, and Ft. Knox. After each administration we examined the psychometric characteristics of the items and scales and revised them for each subsequent administration.

The last administration was at Ft. Knox where about 275 soldiers completed the ABLE and AVOICE. We evaluated the scales for internal consistency, test-retest reliability, and their unique contribution to the predictor battery. For the ABLE scales, the median internal consistency was .84, with a range of .70 to .87. For the AVOICE, the median was .86, with a range of .68 to .96. About 125 soldiers returned two weeks later to complete the ABLE and AVOICE a second time. The median test-retest coefficient for the ABLE was .79, with a range of .68 to .83. For the AVOICE, the median test-retest was .76, with a range of .56 to .86. Uniqueness analyses we conducted show that both the ABLE and AVOICE share very little variance with the ASVAB or with the cognitive and psychomotor tests included in the predictor battery. In short, the psychomotor characteristics of both the ABLE and AVOICE are very good; they are internally consistent, stable over time, and likely to contribute incremental validity to the predictor battery.

Faking Study. The next issue we addressed was faking. The concern was that self-report measures are susceptible to intentional distortion. We, therefore, conducted a faking study, the purpose of which was 1) to determine the extent to which soldiers can distort their responses to temperament and interest inventories when instructed to do so; 2) to determine the extent to which the ABLE response validity scales detect intentional distortion; 3) to determine the extent ABLE response validity scales can be used to adjust or correct scores for intentional distortion; and 4) to determine the extent to which distortion is a problem in an applicant setting.

We gathered data from 125 Army applicants - people who wanted to be accepted into the Army and would have a motive for distorting their responses; we used the Ft. Knox data as an honest comparison sample; and we conducted an experiment in which soldiers were instructed to respond honestly or to distort their responses in a specified way.

The participants in the experimental group were 245 enlisted soldiers at Ft. Bragg. We created four faking conditions: fake good on the ABLE, fake bad on the ABLE, fake interest in combat activities on the AVOICE, and fake interest in non-combat activities on the AVOICE. We also created two honest conditions: honest on the ABLE, and honest on the AVOICE.

The design was a repeated measures with faking and honest conditions counter-balanced. Thus, approximately half the experimental group, or 124 soldiers, completed the inventories honestly in the morning and faked in the afternoon, while the other half (121 soldiers) completed the inventories honestly in the afternoon and faked in the morning. In summary then, we had a 2 x 2 x 2 fixed-factor, completely crossed experimental design.

We performed a multivariate analysis of variance on the ABLE and AVOICE scales separately. All the relevant fake x set interactions for the ABLE were significant at the .01 level, indicating that soldiers can distort their responses. The fake x set x order interactions, significant at the .05 level, indicate that the order in which the conditions occurred has a significant effect on scores. We performed a multivariate analysis of variance on the AVOICE scales and found similar results; people can distort their responses to an interest inventory.

Another research question was the extent to which the response validity scales detected intentional distortion. The results indicate that the Social Desirability scale detects faking good; the effect size of the difference between the means for the honest and fake good conditions is 1.02, or one standard deviation. The Poor Impression scale detects faking bad; the effect size of the difference between the means for the honest and fake bad conditions is 2.67, or just over two and one-half standard deviations.

We next examined the extent to which we could use the response validity scales, Social Desirability and Poor Impression, to adjust ABLE content scales and AVOICE occupational scales for faking. We regressed out Social Desirability from the fake good condition and Poor Impression from the fake bad condition. Table 2 shows the median effect sizes between the honest and faking conditions for the ABLE and AVOICE scales before and after regressing out Social Desirability and Poor Impression. The median difference in ABLE scores between the honest and fake good condition before regressing out Social Desirability is .49 or half a standard deviation. That is, ABLE scale scores differ by about half a standard deviation in the fake good condition as compared to the honest condition. After regressing out Social Desirability from the fake good condition, the ABLE content scales are only .14, or just over 1/10 of a standard deviation, different from the honest condition.

The median difference in ABLE scores between honest and fake bad before regressing out Poor Impression for is 2.10. That is, ABLE content scale scores in the fake bad condition differ by approximately two standard deviations from ABLE content scales in the honest condition. However, after regressing out Poor Impression from the scales, the difference is less than half a standard deviation. Clearly, the response validity scales Social Desirability and Poor Impression can be used to adjust scale scores for the ABLE for intentional distortion. We do not know, however, whether the adjustment formula will cross-validate and be as effective in another data set. Nor do we know whether adjusting the scale scores improves the criterion-related validity of the scales. It may be that the unadjusted scale scores are more criterion-valid than adjusted scores.

We performed the same computations for the AVOICE occupational scales and

Table 2

Effects of Regressing Out Response Validity Scales
(Social Desirability and Poor Impression)
in Faking Conditions for ABLE and AVOICE

	Honest vs Fake Good/Combat Effect Size		Honest vs Fake Bad/Non-Combat Effect Size	
	Before Adjustment	After Adjustment	Before Adjustment	After Adjustment
ABLE Content Scales	.49	.14	2.10	.45
AVOICE Combat Scales	.43	.33	.97	.86
AVOICE Combat-Support Scales	.55	.39	.49	.34

Median values are reported.

found that the results are not nearly as impressive. The bottom two rows show the median effect size of the differences between the honest and faking conditions before and after regressing out the appropriate response validity scale for the AVOICE.

These data demonstrate that: 1) people can distort their responses to temperament and interest scales, 2) response validity scales detect such distortion, and 3) the response validity scales can be used to adjust temperament scale scores for distortion. However, the question remains: To what extent do applicants distort their responses? To answer this question we compared scale scores from the Ft. Bragg experimental honest condition and the Ft. Knox honest condition with the scale scores of approximately 120 Army applicants. These comparisons suggest that applicants do not appear to distort their responses. As shown in Table 3, the applicant means on the temperament scales (ABLE content scales) are lower than one or both of the honest means nine out of eleven times. The results for the AVOICE are similar. In short, applicants do not tent to distort their responses.

Summary. To briefly summarize our approach and results: we identified constructs and developed measures of constructs that had demonstrated criterion-related validity in the past and were judged by expects as likely to be criterion-valid for important Army criteria. The measures we developed contributed unique variance to the predictor battery, were internally consistent or homogeneous, and yielded reliable and stable scale scores across time and motivational conditions.

Our next step is to criterion-validate these measures with Army criteria. Data gathering for that is currently underway.

Table 3

Comparison of Ft. Bragg Honest*, Ft. Knox, and MEPS (Applicants) ABLE Scales

	Ft. Bragg (Honest)*		MEPS (Applicants)		Ft. Knox		Total
ABLE Scale	N	Mean	N	Mean	N	Mean	S.D.
Response Validity Scales							
Social Desirability (Unlikely Virtues)	116	15.91	121	16.63	276	16.60	3.21
Self-Knowledge	116	29.54	121	28.03	276	29.64	3.63
Non-Random Response	116	7.58	121	7.79	276	7.75	.64
Poor Impression	116	1.50	121	1.03	276	1.54	1.84
Content Scales							
Emotional Stability	112	66.22	118	66.03	272	65.05	7.86
Self-Esteem	112	34.77	118	34.04	272	35.12	5.00
Cooperativeness	112	53.33	118	54.60	272	54.19	6.05
Conscientiousness	112	46.37	118	46.49	272	48.97	5.86
Non-Delinquency	112	53.24	118	54.36	272	55.49	6.91
Traditional Values	112	36.67	118	36.97	272	37.28	4.50
Work Orientation	112	59.71	118	58.37	272	61.40	7.73
Internal Control	112	49.48	118	51.90	272	50.37	6.13
Energy Level	112	57.56	118	56.67	272	57.19	6.95
Dominance (Leadership)	112	35.54	118	32.84	272	35.41	6.05
Physical Condition	112	32.96	118	28.27	272	31.08	7.49

*Scores are based on persons who responded to the honest condition first.

References

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Guion, R. M., & Gottier. (1965). Validity of personality measures in personnel section. Personnel Psychology, 18, 135-164.

Tupes, E. C., & Christal, R. E. (May, 1961). Recurrent personality factors based on trait ratings (ASD-TR-61-97). Lackland Air Force Base, TX: Aeronautical Systems Division, Personnel Laboratory.

COMPUTERIZED ASSESSMENT OF PERCEPTUAL AND PSYCHOMOTOR ABILITIES

Jeffrey J. McHenry and Jody L. Toquam

Personnel Decisions Research Institute

Presented on symposium,
"Predicting a Broad Variety of Criteria:
Elaborating the Predictor Space"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Computerized Assessment of Perceptual and Psychomotor Abilities

Jeffrey J. McHenry and Jody L. Toquam

Personnel Decisions Research Institute

One of the main goals of the Army Research Institute's (ARI's) Project A is to develop new predictor measures to supplement the Armed Services Vocational Aptitude Battery (ASVAB). In this paper, we describe 10 new computerized perceptual and psychomotor predictor tests that were pilot tested last fall and are currently being validated in a large-scale concurrent validation study.

The Computer Battery

Toquam, Dunnette, Corpe, and Houston, (1985) have described the procedures used to identify target constructs for cognitive-perceptual predictor test development, and to determine which of these constructs would be measured via paper-and-pencil tests and which would be measured via computer. Following a similar procedure, members of the Project A research team working in the psychomotor ability domain identified two psychomotor ability constructs for predictor test development. Since measurement of both of these constructs required that subjects be presented with a moving stimulus object, it was decided that all psychomotor tests would be presented on the computer.

In total, computer tests were developed for seven constructs (i.e., five cognitive-perceptual ability constructs and two psychomotor ability constructs). To measure these seven constructs, 10 new computer tests were developed. The constructs and tests are listed in Table 1. As Table 1 shows, two tests each were developed to assess reaction time, perceptual speed and accuracy, and precision/steadiness, while one test each was developed to assess the remaining four constructs. (Complete descriptions of each test are available from the authors upon request.)

TABLE 1

Target Constructs and Computer Tests

Target Construct	Definition	Test(s)
Reaction Time	The ability to detect a simple stimulus quickly	Simple Reaction Time Choice Reaction Time
Perceptual Speed and Accuracy	The ability to compare two stimuli and to determine quickly and accurately whether they are the same or different	Perceptual Speed and Accuracy Target Identification
Memory	The ability to encode and store information, and then retrieve that information quickly and accurately	Short Term Memory
Number Facility	The ability to perform simple numerical operations (e.g., addition, subtraction, multiplication, division) quickly and accurately	Number Memory
Movement Judgment	The ability to judge the movement speed and direction of an object and to determine when (or whether) that object will reach a given point in space	Cannon Shoot
Multilimb Coordination	The ability to coordinate the use of two or more limbs (e.g., two hands, two feet, a hand and a foot, etc.) to perform a task	Target Tracking 2
Steadiness/ Precision	The ability to make fine coordinated movements in response to a moving stimulus object	Target Tracking 1 Target Shoot

Pilot Testing

During test development, several pilot tests of portions of the computer battery were conducted at Ft. Carson, Ft. Lewis, and the Minneapolis Military Enlistment Processing Station. A more extensive pilot test of the entire battery was then conducted last fall at Ft. Knox.

The purpose of the pilot testing was to ensure that the tests satisfied three criteria for administration in the Project A concurrent validation study. First, we wanted to ensure that the 10 tests were reliable. Second, we wanted to make certain that the tests did not overlap greatly with the ASVAB. Finally, we wanted to ensure that the computer tests themselves are not highly intercorrelated, since our goal is to measure seven distinct ability constructs with these 10 tests.

Method

Subjects

Subjects included 256 first-term Army enlisted personnel stationed at Ft. Knox. Subjects were drawn from a wide range of MOS. All subjects had been in the service between one and two years at the time of testing.

Procedure

When subjects arrived in the computer testing room, they were asked to take a seat at a testing station. They were told that the computer tests were self-administering so they could work at their own pace. They were instructed to read the instructions carefully, ask questions if they encountered any problems, and try their hardest.

Two weeks later, 121 of the subjects returned for retesting. They were given the same instructions that they had received two weeks earlier and asked to complete the entire computer battery a second time.

Results

Scoring

Responses on computer tests may be used to compute numerous scores. For example, responses to Perceptual Speed and Accuracy items, may be summarized using average decision time, average movement time and average total response time across all items or across only those items in which the subject responds correctly. The average response for each of these may consist of the mean, the median or a trimmed mean computed by deleting the fastest and slowest response times. Other dependent measures derived from this test include the slope and intercept which are computed by regressing the subject's response time against some specified item parameter such as item length. Finally, percent correct can be used as a dependent measure for each subject.

In total, for the 10 tests, 168 different test scores were computed. Preliminary analyses of the reliability of each score and the intercorrelations among the various scores within each test were used to reduce this list to 19 test scores (see Table 2). These 19 scores received more extensive analyses.

Reliability

Table 2 contains the split-half and test-retest reliability for each test score. The majority of split-half reliabilities exceeded .80, and only two are less than .70. As expected, the test-retest reliabilities are lower than the split-half reliabilities. Five test scores have test-retest reliabilities less than .55. In general, those test scores with low test-retest

TABLE 2

Characteristics of the 19 Computer Test Scores

	Reliability		Overlap with ASVAB	
Test Score	r_{sh}	r_{tt}	SMC	Uniqueness
COGNITIVE-PERCEPTUAL TESTS				
Simple Reaction Time - Mean Rt	.90	.37	.07	.83
Choice Reaction Time - Mean Rt	.89	.56	.09	.80
Perc Speed & Acc - Pct Correct	.83	.59	.14	.69
Perc Speed & Acc - Mean RT	.96	.65	.06	.90
Perc Speed & Acc - Slope	.88	.67	.09	.79
Perc Speed & Acc - Intercept	.74	.55	.11	.63
Target Ident - Pct Correct	.84	.19	.05	.79
Target Ident - Mean RT	.96	.67	.16	.80
Short Term Memory - Pct Correct	.72	.34	.10	.62
Short Term Memory - Mean RT	.94	.78	.06	.88
Short Term Memory - Slope	.52	.47	.01	.51
Short Term Memory - Intercept	.84	.74	.11	.73
Number Memory - Pct Correct	.63	.53	.40	.23
Number Memory - Mean Oper RT	.95	.88	.33	.62
Cannon Shoot - Time Score	.88	.66	.02	.86
PSYCHOMOTOR TESTS				
Target Tracking 1 - Mean Log Dist	.97	.68	.23	.74
Target Tracking 1 - Mean Log Dist	.97	.77	.17	.80
Target Shoot - Mean Time to Fire	.91	.48	.06	.85
Target Shoot - Mean Log Dist	.86	.58	.11	.75

reliability are percent correct scores or scores with low split-half reliability.

Overlap with the ASVAB

The squared multiple correlation (SMC) between each test score and the 10 ASVAB subtests is also displayed in Table 2. These SMCs have been adjusted for shrinkage. Only for one test, Number Memory, does the SMC exceed .25. The median SMC across all 19 test scores is .10.

Table 2 also shows the uniqueness for each test score. This value represents an index of the unique (i.e, uncorrelated with the ASVAB) reliable variance of each test score. It is computed by subtracting the SMC with the ASVAB from the split-half reliability. All but two of the uniquenesses in Table 2 exceed .60. This information indicates that these 10 tests have much unique, reliable variance that may contribute to the prediction of job performance.

Overlap among the Computer Tests

Table 3 contains the intercorrelations among the 19 computer test scores. Well over half the intercorrelations between scores on different tests are less than .25, indicating that the various tests are measuring several different abilities.

To determine how we had fared in measuring our target constructs, a principal axis factor analysis was executed. Variables included 17 of the computer test scores (two variables, Perceptual Speed & Accuracy Mean RT and Short Term Memory Mean RT were withheld from the analysis since they correlated .82 and .83 with Perceptual Speed & Accuracy Slope and Short Term

TABLE 3

Intercorrelations among the ASVAB and Pilot Trial Battery (PTB) Tests
Ft. Knox Sample (N=168)

		SRT-RT	CRT-RT	PS&A-PC	PS&A-RT	PS&A-Slp	PS&A-Int	Targ ID-PC	Targ ID-RT	STM-PC	STM-RT	STM-Slp	STM-Int	Can Shoot	No Mem-PC	No Mem-RT	Trk 1-Dist	Trk 2-Dist	TSht-Time	TSht-Dist
PTB -	SRT-RT																			
Computerized	CRT-RT	53																		
Cognitive-	PS&A-PC	17	17																	
Perceptual	PS&A-RT	19	31	50																
Tests	PS&A-Slp	-03	09	52	82															
	PS&A-Int	32	31	-27	-08	-61														
	Targ ID-PC	11	07	40	33	32	-13													
	Targ ID-RT	23	42	20	47	32	13	16												
	STM-PC	-06	04	50	17	26	-23	25	03											
	STM-RT	23	40	26	49	25	23	27	47	08										
	STM-Slp	-06	03	28	29	26	-11	18	13	32	39									
	STM-Int	28	42	10	35	11	31	18	42	-11	83	-19								
	Can Shoot	13	10	00	08	-01	11	-09	25	-02	25	14	18							
	No Mem-PC	-16	-09	29	02	15	-20	14	-12	23	-00	02	-01	-18						
	No Mem-RT	21	24	11	34	21	03	10	27	07	18	15	11	08	-45					
PTB -	Trk 1-Dist	14	25	-12	08	00	11	-04	42	-29	25	-15	35	27	-14	-00				
Computerized	Trk 2-Dist	11	19	-01	11	04	09	02	39	-19	25	-01	27	30	-14	02	81			
Psychomotor	TSht-Time	08	16	16	22	09	12	12	32	09	22	15	15	12	-10	15	23	19		
Tests	TSht-Dist	08	16	-07	03	-00	09	-11	32	-12	27	-16	38	25	-08	02	60	55	-15	

Memory Intercept, respectively), scores from the 10 paper-and-pencil tests described by Toquam et al. (1985), and scores from the 10 ASVAB sub-tests. The sample included only those 168 subjects for whom complete data from all three sets of tests were available. Factor solutions were rotated using the VARIMAX method.

The 7-factor solution was judged the most interpretable. Significant loadings (i.e, greater than .35) for each test score on each factor are shown in Table 4. Based on the factor loadings, we named Factors I-VII general ability, spatial ability, psychomotor ability, general accuracy, basic processing speed, number facility, and a response style factor, respectively. For four of the seven factors (psychomotor ability, general accuracy, basic processing speed, and the response style factor), no paper-and-pencil tests load significantly on these factors. All but one of the tests with significant loadings on the spatial ability factor were paper-and-pencil tests. Both the ASVAB and the computer battery included tests with significant loadings on the other two factors, general ability and number operations; however, the only computer test scores with significant loadings on these factors was Number Memory. Thus, once again, Number Memory appears to be the only computer test that overlaps significantly with the ASVAB.

Some of the factors that include computer tests are moderately similar to the target constructs that we set out to measure with the computer battery. Basic processing speed, for example, contains measures from three target constructs: reaction time, perceptual speed and accuracy, and memory. The number facility factor includes Number Memory test scores, as we had hoped, and also includes the Coding Speed and Number Operations subtests from the ASVAB. Finally, the psychomotor ability factor includes measures of both our target psychomotor ability constructs, multilimb coordination and steadiness/precision.

As Table 4 shows, the time score from Cannon Shoot failed to load significantly on any of the five factors. This indicates that the movement judgment ability tapped by this test differs from the abilities assessed by the other computer tests. This provides indirect evidence that the movement judgment test is measuring a unique perceptual ability, as we had hoped it would.

TABLE 4

Results from a Principal Components Factor Analysis of Scores on the ASVAB, Cognitive Paper-and-Pencil Measures, and Cognitive/Perceptual and Psychomotor Computer Tests[a]
(N = 168)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	h^2
ASVAB GS	75							59
ASVAB AR	75							73
ASVAB WK	77							62
ASVAB PC	62							47
ASVAB NO						84		77
ASVAB CS						62		44
ASVAB AS	62							58
ASVAB MK	77							70
ASVAB MC	63	38	-30					68
ASVAB EI	72							65
Assemb Obj	35	69						66
Obj Rotation		61						49
Shapes		66						51
Mazes		70						67
Path		67	-30					65
Reason 1	37	58						54
Reason 2	37	47						44
Orient 1	37	64						58
Orient 2	40	46			-30			52
Orient 3	60	52						67
SRT-RT					63			44
CRT-RT					61			50
PS&A-PC				67	31			70
PS&A Slope				88				81
PS&A Inter				-65	50			74
Target ID-PC				40				25
Target ID-RT		-41	37		30			57
STM-PC				39			34	41
STM-Slope							41	25
STM-Int			38		51			47
Cannon Shoot			32					19
NM-PC	53					37		52
NM-RT	-37					-46		54
Tracking 1			86					82
Tracking 2			77					66
Target Shoot-TF							42	23
Target Shoot-Dist			64					48
Variance Explained	5.69	4.70	2.83	2.37	1.92	1.87	1.17	

[a] Note that the following variables were not included in this factor analysis: APQT, PS&A Reaction Time and Short Term Memory Reaction Time.

(Please also note that decimals have been omitted.)

Discussion

All of the tests except Simple Reaction Time yielded at least one test score with split-half reliability in excess of .80 and test-retest reliability in excess of .55. Thus, we met our first goal, which was to ensure that all the computer tests attained adequate levels of reliability.

Our second goal was to ensure that the new computer tests were not redundant with the ASVAB. SMCs between the 19 test scores and the ASVAB tended to be quite low. Uniquenesses indicated that the computer tests had the potential to contribute a great deal of unique, reliable variance to the prediction of job performance. Thus, we also met our second goal.

Analyses designed to evaluate the intercorrelations among the new tests showed that the various tests generally shared little common variance. Results from a factor analysis indicate that there were at least five (and probably six) different ability factors underlying performance on the 10 tests; these factors are moderately similar to the target constructs we set out to measure. It is important to note here that results from the factor analysis must be considered tentative at best because the sample size includes only 168 subjects. Data obtained from the ongoing concurrent validity study with over 10,000 subjects will provide us with more stable information about our constructs and the relationships among those constructs.

Generally, we felt that the results of the pilot testing indicated that only minor modifications were required in the tests prior to concurrent validation testing. Our observations of subjects during pilot testing suggested a number of changes in the instructions for virtually all of the tests. The split-half reliability data indicated that several of the tests could be shortened without any significant impact on test reliability. Finally, there was some evidence (not discussed in this paper, but noted in McHenry & McGue, 1985) that the two Target Tracking Tests should be made more difficult and that the Target Shoot Test should be made easier. Aside from these, few modifications were made in the computer battery prior to concurrent validation testing. (See Toquam, Dunnette, Corpe, McHenry, Keyes, McGue, Houston, Russell & Hanson, 1985, for more detailed information regarding changes in the computerized perceptual tests.)

Presently, concurrent validation testing is winding down. By the middle of next month, we will have collected predictor and criterion data on almost 10,000 first-term Army enlisted personnel in 19 MOS. It is our hope that at this time next year, we will be able to present some initial validity data for our 10 new computerized perceptual and psychomotor tests.

References

McHenry, J. J., & McGue, M. K. (1985). *Problems, issues, and results in the development of computerized psychomotor measures.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., McHenry, J. J., Keyes, M. A., McGue, M. K., Houston, J. S., Russell, T. L., & Hanson, M. A. (1985). *Development of cognitive/perceptual measures: Supplementing the ASVAB.* Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Toquam, J. L., Dunnette, M. D., Corpe, V. A., & Houston, J. S. (1985). *Adding to the ASVAB: Cognitive/perceptual measures.* Paper presented at the 27th Annual Military Testing Association Conference, San Diego.

EXAMINATION OF ENVIRONMENTAL DETERMINANTS OF ARMY PERFORMANCE CRITERIA

Darlene M. Olson
U.S. Army Research Institute

Walter C. Borman
Personnel Decisions Research Institute

Presented on panel,
"Organizational Effectiveness"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

EXAMINATION OF ENVIRONMENTAL DETERMINANTS OF ARMY PERFORMANCE CRITERIA

Darlene M. Olson
U.S. Army Research Institute[1]

Walter C. Borman
Personnel Decisions Research Institute

Job performance has been conceptualized as a product of individual attributes, abilities, and skills which are measurable at the time an individual first enters the organization, of environmental/organizational variables which impact on the individual after job-entry and of the person's motivation to perform. Previous empirical research has investigated work performance in terms of taxonomies of human abilities, values, and personality characteristics (Dunnette, 1976). However, until recently little research has focused on developing taxonomies of environmental/organizational variables or examining relationships between these factors and work-related outcomes.

The major purpose of this research was to examine relationships among individual, organizational/environmental factors, job characteristic variables, and measures of both maximal (e.g., hands-on and job knowledge tests) and typical (e.g., supervisory and peer ratings of performance) performance criteria for first-term soldiers in the Army. This paper discusses results from administering a 110-item Army Work Environment Questionnaire (AWEQ) to 800 first-term enlisted personnel from five military occupational specialties (MOS).

A major impetus for research on environmental variables was the work of Schneider (1978), who proposed that such situational influences as job/task characteristics, organizational practices (e.g., reward system) and climate variables could either directly influence performance or moderate the relationship between cognitive abilities and performance. During the early 1980's several research projects were initiated to develop empirically validated taxonomies of environmental variables (e.g., Peters & O'Connor, 1980; Olson, Borman, Roberson, & Rose, 1984). In a series of laboratory studies conducted by Peters and O'Connor, and their colleagues (for a review see Eulberg, O'Connor, Peters & Watson, 1984), results have demonstrated that situational constraints are significantly related to ineffective task performance, job dissatisfaction, and increased frustration.

Although correlational field studies have supported the relationships between environmental/situational variables and affective reactions to the job (e.g., satisfaction), associations between these factors and ratings of performance effectiveness have been inconsistent.

In general, the mixed results found for relationships between environmental factors and performance suggest that the magnitude of the correlation coefficients are dependent on the level of inhibitors/facilitators

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

actually present in the work environment. Further, the ways situational variables are conceptualized, the kinds of jobs investigated, and the types of performance criteria examined may impact on the observed relationships.

METHOD

Subjects. The research sample contained 800 first-term enlisted personnel from five Army jobs. There were 172 infantrymen (11B MOS), 169 armor crewmen (19E MOS), 144 radio teletype operators (31C MOS), 155 light wheel vehicle mechanics (63B MOS), and 160 medical care specialists (91A MOS). These MOS were sampled at four continental United States and two European Army installations.

Measures. An assessment battery containing an environmental questionnaire and a comprehensive set of typical (e.g., supervisory ratings) and maximal (e.g., job knowledge test) performance measures was used in this research.

Army Work Environment Questionnaire (AWEQ). The Army Work Environment Questionnaire is a 110-item multiple choice instrument that measures 14 dimensions of the Army work environment. The AWEQ was constructed in a two-stage process (Olson, et al. 1984). Briefly, in Stage I, a taxonomy of first-tour environmental influences on soldier performance was derived through application of a critical incident methodology. A total of 282 critical incidents, generated by Army experts ($N = 67$) and independently content-analyzed by six psychologists, identified environmental/organizational influences beyond the control of the soldier that had a significant impact on performance, either inhibiting or facilitating that performance. The Army work environment taxonomy contains the following nine "job-oriented" factors: (1) Resources/tools/equipment, (2) Workload/Time Availability, (3) Training, (4) Physical Working Conditions, (5) Job-Relevant Information, (6) Job Relevant Authority, (7) Perceived Job Importance, (8) Work Assignment, and (9) Changes in Job Procedures/ Equipment, as well as, the remaining, five "climate-oriented" dimensions: (10) Reward System, (11) Discipline, (12) Individual Support, (13) Job Support/Guidance and (14) Role Models. In Stage II, items were written to cover the content of the 14 environmental dimensions.

Items on the AWEQ are descriptive in nature and respondents are asked to indicate on a 5-point rating scale (e.g., 1 = Very Seldom or Never to 5 = Very Often or Always) how often each environmental situation described in the items occurs on their present job.

Job Performance Measures. The set of typical and maximal performance criteria used in this study was developed as a component of a broader research program conducted under Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel. This comprehensive nine year research effort was initiated to help the Army access, assign, and retain quality personnel.

The typical performance criteria included supervisory and peer job performance ratings. Separate behaviorally-anchored rating scales (BARS), derived from a critical incident job analysis procedure, were used to measure both the MOS (job)-specific and Army-wide components of soldier performance and effectiveness on a 7-point behavior rating format. For each research participant in the five MOS, an Army-wide and MOS-specific rating was computed by averaging the performance ratings across all individual dimensions for supervisors and peers separately.

The maximal performance criteria included hands-on (work sample) tests and job knowledge measures. The hands-on tests for each MOS consisted of 15 tasks identified for the MOS. The individual performance components of each task were scored by trained raters on a pass-fail basis and an overall hands-on score was computed for each soldier by averaging the proportions passed across the tasks tested. Multiple-choice tests were developed to assess job knowledge relevant to each important task for an MOS. An overall job knowledge test score for each research participant was derived as a percentage of the number of items answered correctly.

Procedures. After the supervisor and peer raters were trained to use the Army-wide and MOS-specific BARS, they evaluated the job performance of soldiers in the research sample. Concurrently with these assessments, first-tour soldiers participating in the research were administered: (a) the Army Work Environment Questionnaire and (b) the appropriate job knowledge and hands-on test. For all respondents, scores on the environmental measure were merged with scores from the maximal and typical performance criteria for analyses.

RESULTS AND DISCUSSION

For the total sample, Table 1 presents the means, standard deviations, and reliability coefficients for the research measures. When mean ratings on the AWEQ scale dimensions are collapsed across MOS and installation, results suggest that a complex set of both facilitating and inhibiting influences describe the Army work environment. For instance, the mean ratings for such AWEQ scales as Training ($\underline{M}$ = -3.02), Work Assignment ($\underline{M}$ = -1.90), Reward System ($\underline{M}$ = -1.75), and Job Support ($\underline{M}$ = -1.42) were described somewhat negatively. In contrast, such environmental variables as Perceived Job Importance ($\underline{M}$ = 1.76), Discipline Practices ($\underline{M}$ = 1.10), Individual Support ($\underline{M}$ = .79), and adequacy of Role Models ($\underline{M}$ = .74) were generally described more positively. Uncorrected reliability estimates displayed in Table 1 show that the job knowledge tests tend to be the most reliable of the maximal performance criteria and the Army-wide BARS (supervisors) have the largest coefficients of the typical performance measures. Generally, the AWEQ scale scores, with coefficients ranging from .57 to .78, have adequate reliabilities for a research instrument.

Table 2 presents the intercorrelation matrix for the AWEQ scales. Intercorrelations among the 14 AWEQ scales show that the climate-oriented dimensions are more highly related than the job-oriented factors.Subsequent test development work on the AWEQ, which has included an item-analysis and a principle component factor analysis with a varimax rotation, has been conducted to identify a subset of the original 110 items that best define the factor structure of the AWEQ. Although findings from these analyses corroborate the redundancy displayed in Table 2 for some of the AWEQ scales and tentatively suggest that a five factor solution with 53 items may permit a more parsimonious explanation of the underlying Army work environment constructs, results based on the revised-AWEQ have not been sufficiently cross-validated. Hence, results presented in Table 3 focus on the relationships between the 14 scale scores from the conceptual taxonomy, and a comprehensive set of both ratings of job performance and more objective performance indices.

Table 1

Means, Standard Deviations, and Reliability Coefficients for Selected Measures Across MOS.

Measures	*N*	*M*	*SD*	*r*[1]
Army-Wide BARS (Peers)	727	4.52	.71	.78-.86
Army-Wide BARS (Supervisors)	722	4.50	.84	.81-.86
MOS-Specific BARS (Peers)	727	4.60	.66	.76-.86
MOS-Specific BARS (Supervisors)	718	4.62	.77	.78-.87
Hands-on Test	685	71.72	16.11	.35-.56
Job Knowledge Test	745	62.47	10.63	.84-.91
AVEQ Scales (N of items):[2]				
Resources (n=7)	734	-.99	4.96	.75
Workload (n=8)	752	-.67	4.34	.58
Training (n=11)	736	-3.02	5.91	.64
Physical Working Conditions (n=6)	741	.67	3.83	.57
Job Authority (n=6)	760	-.25	3.65	.57
Job Information (n=8)	726	.45	4.60	.67
Job Importance (n=7)	725	1.76	4.65	.67
Work Assignment (n=9)	731	-1.90	6.80	.70
Changes in Job Procedures (n=8)	745	-.89	4.21	.58
Reward System (n=7)	736	-1.75	5.14	.78
Discipline (n=6)	751	1.10	4.07	.65
Individual/Support (n=9)	727	.79	5.46	.73
Job Support (n=8)	734	-1.42	5.12	.72
Role Models (n=10)	731	.74	5.98	.71

Note. 1). For performance ratings, the range of interrater reliabilities across MOS are reported.

For Hands-on and Job Knowledge tests, the range of split-half reliabilities across MOS are reported.

For the Environmental scales, Cronbach's alpha coefficients are used as measures of internal consistency.

2). Mean scale scores were computed such that "0" is a neutral environment. Positive mean values indicate positive descriptions of the environment for that scale. Negative scale means indicate the opposite.

Table 2

Scale Intercorrelations for the AVEQ.

AVEQ Scales*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Resources	-													
2. Workload	.52	-												
3. Training	.29	.26	-											
4. Working Conditions	.55	.48	.23	-										
5. Job Authority	.47	.52	.39	.51	-									
6. Job Information	.52	.50	.42	.50	.60	-								
7. Job Importance	.21	.20	.30	.22	.32	.33	-							
8. Work Assignment	.26	.24	.66	.18	.35	.36	.43	-						
9. Job Procedures	.49	.51	.44	.43	.50	.51	.24	.40	-					
10. Reward System	.38	.40	.40	.37	.58	.56	.29	.33	.45	-				
11. Discipline	.31	.31	.18	.37	.47	.48	.30	.14	.36	.45	-			
12. Individual Support	.31	.32	.35	.36	.56	.60	.34	.27	.41	.62	.54	-		
13. Job Support	.39	.40	.44	.38	.64	.62	.34	.37	.48	.73	.48	.72	-	
14. Role Models	.41	.46	.44	.42	.61	.60	.34	.35	.50	.56	.48	.58	.65	-

Note. All AVEQ scale intercorrelations are significant at $p < .05$.

*Correlations significant at $p < .05$.

1). Scales 1-9 are more job-oriented and scales 10-14 are more climate-oriented.

Table 3 presents the correlation coefficients between the 14 AWEQ scale scores and the set of performance criteria for the total sample. Several interesting findings emerged. First, the largest correlations were found between environmental variables and typical performance measures, specifically the Army-wide BARS. In terms of the number of significant effects, 46.4% of the correlation coefficients between environmental variables and typical measures, as compared with 28.6% of the correlations for maximal criteria, were statistically significant. This difference cannot be attributed to sampling error, since differences in sample sizes for the correlational values shown in Table 3 were relatively minor.

Second, generally the environmental dimensions of (a) Perceived Job Importance, (b) Discipline practices, (c) Individual Support, and (d) the Reward System tended to be significantly correlated with performance criteria for the total sample. In contrast, the AWEQ scale scores on (a) Resources/Tools/Equipment, (b) Workload/Time Availability, (c) Physical Working Conditions, and (d) Changes in Job Procedures/Equipment were not significantly associated with scores on the performance measures. Although the magnitude of these environment-performance relationships are lower than those previously reported with Army field test data from Project A (see Olson et al., 1984), fairly consistent trends have been observed in the pattern of significant relationships between climated-oriented AWEQ scales and performance ratings.

Table 3

Correlations Between AWEQ Scale Scores and Performance Criteria.

Performance Criteria	Scale Scores on Army Work Environment Questionnaire													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
							Typical Performance Measures							
Army-wide BARS (Peers)	.05	.02	.07	.08*	.13*	.09*	.23*	.07	.04	.11*	.14*	.18*	.14*	.11*
Army-wide BARS (Supervisors)	.01	-.03	.08*	.06	.15*	.11*	.17*	.12*	.01	.11*	.13*	.14*	.13*	.09*
MOS-specific BARS (Peers)	.01	-.04	.06	.05	.07	.04	.16*	.04	0	.05	.10*	.09*	.07	.05
MOS-specific BARS (Supervisors)	-.01	-.08*	.08*	.03	.06	.06	.13*	.11*	-.01	.03	.06	.05	.03	.03
							Maximal Performance Measures							
Hands-on Test	-.02	0	-.06	-.01	-.04	.02	.09*	-.02	-.02	-.08*	.08*	.04	-.07	.04
Job Knowledge Test	-.05	-.05	-.05	.03	0	.03	.13*	-.07	-.08*	-.09*	.13*	.11*	-.03	.01

Note. AWEQ SCALES: 1= Resources, 2=Workload, 3=Training, 4=Physical Working Conditions, 5=Job Relevant Authority, 6=Job Relevant Information, 7=Perceived Job Importance, 8=Work Assignment, 9=Changes in Job Procedures, 10=Reward System, 11=Discipline, 12=Individual Support, 13=Job-Related Support, 14=Role Models.

*Correlations which are significant at $p < .05$.

Third, when relationships between typical performance measures and environmental factors were examined, 60% of the correlations between climated-related dimensions and 38.9% of the correlations with job-oriented factors were significantly related to performance ratings. Further, a similar pattern of significant relationships was found between the environmental variables and maximal performance criteria, specifically 50% of the observed correlation coefficients for climate dimensions and 16.7% of the correlations for job dimensions were significantly associated with scores on maximal performance measures. It was predicted that job-oriented environmental factors should have more significant relationships with the objective, maximal performance measures, than the supervisory and peer ratings of overall soldier effectiveness. However, these findings did not support this contention, because a larger percentage of climate-oriented factors than job-oriented factors were significantly correlated with both types of performance indices.

Finally, consistent relationships were observed between environmental variables and the typical performance measures, specifically the Army-wide BARS, regardless of whether performance was evaluated by supervisors or peers. This finding indicates the existence of some convergence across types of performance criteria with respect to the influence of environmental factors.

CONCLUSIONS

This research examined correlations between 14 scale scores on an Army Work Environment Questionnaire and measures of both typical and maximal performance. Prior to this applied research in an Army setting, inconsistent findings were reported in the empirical literature with respect to relationships between organizational/environmental variables and performance.

Results from this applied Army research indicated that significant relationships exist between job-oriented and climate-related environmental variables and both job performance ratings (typical measures) and more maximal, objective criteria-job knowledge and hands-on tests. Further, these findings suggest that: (1) environmental factors have their strongest correlations with more typical performance measures such as Army-wide BARS and (2) climate-oriented environmental variables have a larger number of significant effects on maximal performance criteria than job-related environmental dimensions. Perhaps, the weak but significant correlations observed between environmental dimensions and performance may be related to: (1) a lack of sufficiently constraining or facilitating conditions on the part of the environmental variables themselves or (2) contextual factors such as raters adjusting their performance evaluations to compensate for the negative/positive effects of specific work environments.

REFERENCES

Dunnette, M. D. (Ed.) (1976). Handbook of industrial and organizational psychology. Chicago, IL.: Rand McNally.

Eulberg, J. R., O'Connor, E. J., Peters, L. H., & Watson, T. W. (1984). Performance constraints: A selective review of relevant literature. Psychological Documents.

Olson, D. M., Borman, W. C., Roberson, L., & Rose, S. R. (1984). Relationship between scales on an Army work environment questionnaire and measures of performance. Paper presented at the 92nd annual meeting of the American Psychological Association, Toronto, Canada.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influence of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Schneider, B. (1978). Person-situation selection: A review of some ability-situation interaction research. Personnel Psychology, 31, 381-397.

MAPPING PREDICTORS TO CRITERION SPACE: OVERVIEW

Norman G. Peterson
Personnel Decisions Research Institute

Presented on symposium,
"Predicting a Broad Variety of Criteria:
Elaborating the Predictor Space"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Mapping Predictors to Criterion Space: Overview

Norman G. Peterson
Personnel Decisions Research Institute

Introduction

Our applied problem is to expand the presently measured predictor space for the ultimate purpose of accurately selecting persons for the U.S. Army and appropriately classifying those persons into jobs or Military Occupational Specialities (MOS). In this paper, I describe the strategy we have adopted, the thinking behind the strategy, and some of the progress that has been made following our strategy. A fuller description can be found in Peterson, 1985.

As you all know, the U.S. Army presently has a lot of jobs and hires, almost exclusively, inexperienced and untrained persons to fill those jobs. One implication of these obvious facts is that a highly varied set of individual differences' variables must be put into use to stand a reasonable chance of improving the present level of accuracy of predicting training performance, job performance, and attrition/retention in a substantial proportion, if not all, of those jobs. Much less obvious is the particular content of that set of individual differences variables, and the way the set should be developed and organized; or put another way, how the predictors should be mapped onto the criterion space.

Theoretical Approach

We have approached this problem by adopting a construct-oriented strategy of predictor development, but with a healthy leavening from the content-oriented strategy. Essentially, we endeavored to build up a model of predictor space by (a) identifying the major, relatively independent domains or types of individual differences' constructs that existed; (b) selecting measures of constructs within each domain that met a number of psychometric and pragmatic criteria, and (c) further selecting those constructs that appeared to be the "best bets" for incrementing (over present predictors) the prediction of the set of criteria of concern (i.e., training/job performance and attrition/retention in Army jobs). Ideally, the model would, we hoped, lead to the selection of a finite set of relatively independent predictor constructs that were also relatively independent of present predictors and maximally related to the criteria of interest. If these conditions were met, then the resulting set of measures would predict all or most of the criteria, yet possess enough heterogeneity to yield powerful, efficient classification of persons into different jobs. The development of such a model also had the virtue that it could be at least partially "tested" at many points during the research effort, and not just at the end, when all the predictor and criterion data are in. For example, we could examine the covariance of newly developed measures with one another and with the present predictors, notably the ASVAB. If the new measures were not relatively independent of ASVAB and measures from other domains as predicted by the model, then we could take steps to correct that. Also, by constructing such a visible model, we thought that modifications and improvements could be much more straightforwardly implemented.

Figure 1 presents an illustrative, construct-oriented model and is presented in order to represent the model in abstract. Note that both the criterion and predictor space are depicted. A great deal of the work of Project A is devoted to describing and defining the job performance criterion and we, on the predictor side, have made much use of the information coming from those efforts.

If this illustrative model were to be developed and tested with data, then the network of relationships on the predictor side, the criterion side, and between the two could be confirmed, disconfirmed, and/or modified. It goes without saying, but I will say it anyway, that the development of such models must be done very carefully and conservatively, and subjected frequently to reality testing. We have kept this firmly in mind, Note, however, that the possession of such a model enables one to state fairly clearly why such a predictor is being researched, and to check quickly, at least rationally, whether or not the addition of a predictor is likely to improve prediction.

Finally, the model is depicted as a matrix with a hierarchical arrangement of both the rows and columns. We have found it very useful to employ this hierarchical notion, since it allows us to think in terms of appropriate levels of specificity for a particular problem as we do the research, or for future applications of measures.

We began our research with a general kind of model, very much like the one presented in Peterson and Bownas (1982). That is, we conceived of the predictor space as divided into several domains with major, relatively independent constructs falling into each domain. At this early point in the research, we were most concerned with thinking about the predictor space in a way guided by past research that would also provide "handles," if you will, for us to approach our particular applied problem. We formed

		Criteria							
		Training Performance			Job Task Performance		Attrition/ Retention		
Predictors		Pass/ Fail	Test Grades	Atten-dance	Common Tasks	Specific Tasks	Finish Term	Reen-list	Early Discharge
Cognitive	Verbal	H*	H	L	M	M	L	L	L
	Numerical	M	H	.	.	.	.		
	Spatial								
Psychomotor	Precision								
	Coordination								
	Dexterity								
Temperament	Dependability								
	Dominance								
	Sociability								
Interests	Realistic								
	Artistic								
	Social	.	.	.	M	M	M	L	L

FIGURE 1. Illustrative Construct-Oriented Model

*Denotes expected strength of relationship, High, Medium, Low.

three domain teams to be responsible for broad pieces of this predictor space model, to wit: a "non-cognitive" team for temperament, biographical data, and vocational interest variables; a "cognitive" team for cognitive and perceptual variables; and a "psychomotor" team for psychomotor variables.

Literature Review. The domain teams began with a large-scale literature review. Within each area, the teams carried out essentially the same steps. These were: 1) compile an exhaustive list of possibly relevant reports, articles, books or other sources; 2) review each source and determine its relevancy for the project by examining the title and abstract (or other brief review); 3) obtain the sources identified as relevant in the second step; and 4) for relevant materials, carry out a thorough review and transfer relevant information onto special review forms developed for the project.

Within the first step, several activities were carried out to insure as comprehensive a list as possible. Several computerized searches of relevant data bases were done. In addition to the computerized searches, we obtained reference lists from recognized experts in each of the areas, emphasizing the most recent research in the field. We also obtained several annotated bibliographies from military research laboratories. Finally, we scanned the last several years' editions of research journals that are frequently used in each ability area as well as more general sources such as textbooks, handbooks, and appropriate chapters in the Annual Review of Psychology.

The vast majority of the sources identified as described above were not relevant to our purpose. These non-relevant sources were weeded out in Step 2. After obtaining the relevant sources, these were reviewed and two forms were completed for each source: an Article Review form and a Predictor Review form (several of the latter form could be completed for each source.) These forms were designed to capture, in a standard format, the essential information from the reviewed sources, which varied considerably in their organization and reporting styles. The output of the literature search, in the form of the completed review forms and copies of the actual sources, served as input to several later steps.

Expert Judgments. One of these steps was the identification of a set of predictor constructs that met a number of psychometric and practical criteria. There were twelve such criteria used to evaluate constructs, like reliability, criterion-related validity, robustness and ease of administration procedures, etc. At least two researchers evaluated each construct on these twelve factors, using five point scales, and these evaluations guided the selection of 53 predictor constructs.

Definitions of these selected constructs were written and descriptive materials (psychometric data, validity evidence, and illustrative items) were prepared. These materials were used in an expert judgment process wherein 35 experienced personnel and research psychologists estimated the "true validity" of each of the 53 predictor constructs for each of 72 Army enlisted criteria. These 72 criterion descriptions were prepared by Project A researchers who were focusing on describing the job performance of Army enlisted ranks. (See Wing, Peterson, and Hoffman, 1984, for a complete description of this expert judgment process.)

These expert judgments proved to be highly reliable (the reliability of the pooled raters' estimates of validity of each construct for each criterion was over .90), and factor analysis of their ratings provided our first model of the predictor space. Figure 2 shows that model. This

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension	A. Verbal Ability/ General Intelligence	COGNITIVE ABILITIES
5. Reading Comprehension		
16. Ideational Fluency		
18. Analogical Reasoning		
21. Omnibus Intelligence/Aptitude		
22. Word Fluency		
4. Word Problems	B. Reasoning	
8. Inductive Reasoning: Concept Formation		
10. Deductive Logic		
2. Numerical Computation	C. Number Ability	
3. Use of Formula/Number Problems		
12. Perceptual Speed and Accuracy	H. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory	J. Memory	
17. Follow Directions		
19. Figural Reasoning	F. Closure	
23. Verbal and Figural Closure		
6. Two-dimensional Mental Rotation	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
7. Three-dimensional Mental Rotation		
9. Spatial Visualization		
11. Field Dependence (Negative)		
15. Place Memory (Visual Memory)		
20. Spatial Scanning		
24. Processing Efficiency	G. Mental Information Processing	INFORMATION PROCESSING
25. Selective Attention		
26. Time Sharing		
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests	M. Realistic vs. Artistic Interests	
51. Artistic Interests (Negative)		
28. Control Precision	I. Steadiness/Precision	PSYCHOMOTOR
29. Rate Control		
32. Arm-hand Steadiness		
34. Aiming		
27. Multilimb Coordination	D. Coordination	
35. Speed of Arm Movement		
30. Manual Dexterity	K. Dexterity	
31. Finger Dexterity		
33. Wrist-Finger Speed		
39. Sociability	Q. Sociability	SOCIAL SKILLS
52. Social Interests		
50. Enterprising Interests	R. Enterprising Interest	
36. Involvement in Athletics and Physical Conditioning	T. Athletic Abilities/Energy	VIGOR
37. Energy Level		
41. Dominance	S. Dominance/Self-esteem	
42. Self-esteem		
40. Traditional Values	N. Traditional Values/Conventionality/Non-delinquency	MOTIVATION/ STABILITY
43. Conscientiousness		
46. Non-delinquency		
53. Conventional Interests		
44. Locus of Control	O. Work Orientation/Locus of Control	
47. Work Orientation		
38. Cooperativeness	P. Cooperation/Emotional Stability	
45. Emotional Stability		

FIGURE 2. Hierarchical Map of Predictor Space

PILOT TRIAL BATTERY	CLUSTERS	FACTORS
ASVAB	A. Verbal Ability/ General Intelligence	COGNITIVE ABILITIES
Reasoning 1 and 2	B. Reasoning	
Number Memory (c)	C. Number Ability	
Perceptual Speed and Accuracy (c) Target Identification (c)	H. Perceptual Speed and Accuracy	
AVOICE	U. Investigative Interests	
Short Term Memory (c)	J. Memory	
Reasoning 1 and 2	F. Closure	
Assembling Objects Object Rotation Shapes Mazes Path Orientation 1, 2, and 3	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
Simple Reaction Time (c) Choice Reaction Time (c)	G. Mental Information Processing	INFORMATION PROCESSING
ASVAB	L. Mechanical Comprehension	MECHANICAL
AVOICE	M. Realistic vs. Artistic Interests	
Target Tracking 1 (c) Target Shoot (c)	I. Steadiness/Precision	PSYCHOMOTOR
Target Tracking 2 (c) Target Shoot (c)	D. Coordination	
--	K. Dexterity	
ABLE/AVOICE	Q. Sociability	SOCIAL SKILLS
AVOICE	R. Enterprising Interest	
ABLE	T. Athletic Abilities/Energy	VIGOR
ABLE	S. Dominance/Self-esteem	
ABLE	N. Traditional Values/Conventionality/Non-delinquency	MOTIVATION/ STABILITY
ABLE	O. Work Orientation/Locus of Control	
ABLE	P. Cooperation/Emotional Stability	
Cannon Shoot (c)	Movement Judgment	

(c) = Computerized Measures

FIGURE 3. Pilot Trial Battery Measures of the Modeled Predictor Space

model represents the predictor structure in terms of their covariances with each other based on their judged validity relationships to dimensions of Army enlisted criteria.

Test Construction. Figure 2 served as a blueprint of sorts for our test construction efforts. The three domain teams set about writing tests and inventories to measure the constructs shown there. We went through a fairly extensive process of writing (or, in the case of computerized tests, programming) instruments, trying them out at Army sites (MEPS and/or Army forts), then revising the instruments based on the tryout results. After about four such iterations (at Minneapolis MEPS, Fts. Carson, Campbell, and Lewis), we possessed a set of instruments collectively labeled the Pilot Trial Battery. That set of measures is shown in Figure 3.

Note that the measures are slotted into the cluster and factor space, insuring that we adequately operationalized the model. Note also that one measure, "cannon shoot", is included and it measures Movement Judgment, a variable that was not originally included. It was added because it seemed to be a variable that was important for a variety of combat arms MOS, but had escaped our notice because of a dearth of research on such a variable.

This Pilot Trial Battery consumed approximately six and one-half hours of testing time and the entire battery was administered to a sample of about 250 soldiers at Ft. Knox. Test-retest data were also collected. Analyses of these data were used to further revise the measures and to reduce the battery in size so that it could be administered in four hours. The reduction in the size of the battery was accomplished by deleting some tests entirely and by deleting items from other tests. (The tests deleted were Reasoning 2, Shapes, Path, and Orientation 1.) The existence of the predictor model proved especially helpful to those of us faced with the hard decision of deleting tests and items. The impact of various decisions in terms of coverage of the "predictor space" could readily be seen and, along with the tryout data, empirically evaluated.

This revised and reduced battery was labeled the Trial Battery and is presently being administered to a large sample (N=11,000) of soldiers in the U.S. and Europe in a concurrent validity study. In terms of testing time, 34% of the battery is devoted to the computerized perceptual/psychomotor measures, 50% to cognitive paper-and-pencil measures, and 16% to non-cognitive, paper-and-pencil inventories. Once the concurrent validity data are in hand, we will be able to make some fairly definitive tests of our model--in terms of its factorial structure, validity, and classification efficiency.

References

Peterson, N. G., & Bownas, D. A. (1982). Skill, task structure, and performance acquisition. In Marvin D. Dunnette and Edwin A. Fleishman (Eds.), Human performance and productivity (Vol. I). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Peterson, N. G. (1985). Overall strategy and methods for expanding the measured predictor space. Paper presented at the 93rd Annual Convention of the American Psychological Association, Los Angeles.

Wing, H., Peterson, N. G., & Hoffman, R. E. (1984). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto.

USING MICROCOMPUTERS FOR ASSESSMENT: PRACTICAL PROBLEMS AND SOLUTIONS

Rodney L. Rosse and Norman Peterson
Personnel Decisions Research Institute

Presented on symposium,
"Predicting a Broad Variety of Criteria:
Elaborating the Predictor Space"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Using Microcomputers for Assessment: Practical Problems and Solutions

Rodney L. Rosse and Norman Peterson
Personnel Decisions Research Institute

Introduction

"History repeats itself" is an adage that probably does not apply to the advances of microprocessor developments. Given the frantic rate of development, it is difficult to imagine that circumstances could ever again occur in just the way that they did at the outset of this effort in the Fall of 1983. It would seem, however, that any 1986 project might be enhanced by consideration of both the occasional wisdom and sometime folly of our beginning efforts.

Initially, even the goals to be accomplished were far from obvious and may have remained beyond our vision except for the valuable help obtained through visits to several research centers doing advanced work in computerized testing: (1) Air Force Human Resources Laboratory at Brooks Air Force Base, Texas, (2) Army Research Institute Field Unit at Fort Rucker, Alabama, (3) Naval Aerospace Medical Research Laboratory, Pensacola, Florida, and (4) Army Research Institute Field Unit at Fort Knox, Kentucky. Experimental testing projects using computers at these sites had already produced impressive developments which stimulated the ideas of the project at hand and have continued to influence our work.

In this paper, we focus primarily on the process we followed and some problems we encountered in hardware and software acquisition and development for the purpose of developing new predictor tests of abilities that could best be administered via microprocessors.

Hardware Acquisition and Development

Much of the detail of the planned products was yet to evolve at the point of acquisition of the first six machines so that we had to focus upon more general objectives. It was clear that we wished to accomplish several things which were either difficult or impossible to accomplish with paper-and-pencil testing. Specifically, we required the ability to have a very high degree of precision in stimulus presentation and a high degree of control of respondent behavior. Dependent variables were specifically expected to include precision in timing of stimulus presentation and response speed.

Microprocessor. The choice of which microprocessor to use for the preliminary development was not obvious. The arrays of available microcomputer devices were, at the time, in transition from earlier machines which used the first popular microprocessor chips (i.e., 8080 or Z-80) into a newer variety of options created by the influence of IBM's entry into the market with their "PC" employing the newer 8088, 8086-7 chips. With the newer machines came more flexible operating systems (e.g., DOS 1 or DOS 2).

A computer designed for portable use was deemed to be a highly desirable characteristic because the machines were to be frequently disassembled, carried to new locations, and reassembled by non-technical personnel. Such portable machines had been available only briefly so that little reported experience with them was available.

We acquired six machines made by Compaq (TM) which appeared to suit the need. They were among the "newer" types of machines which used a variation of the MS-DOS operating system. They were equipped with standard game adapters which permitted the analog inputs from "off-the-shelf" joysticks and boolean input from game button switches.

The choice was specifically made to avoid using color in the visual displays for at least two reasons: (1) the certainty of individual differences in color vision among military recruits, and (2) dread of the prospects of attempting to calibrate video colors for standardization of presentation. Accordingly, we precluded the possibility of directly investigating the value of stimulus effects in color presentation.

The graphics capability of the Compaq microcomputer proved to be minimally acceptable for the applications which were to come. In graphics mode, the pixels (or dots) on the screen are orgaııized into 200 rows and 640 columns. More recently, several computers of the "personal" computer type are offering 400 rows with 640 columns which should provide improved resolution.

Very accurate timing of events occurring in the testing process was essential. Initially, timing was accomplished by two means: (1) accessing the calendar clock that is available in any machine which uses MS-DOS (or the variations of MS-DOS that are sold under computer tradenames), and (2) use of calibrated software loops. Without delving too far into technical details, those two options eventually presented some difficulties because of time consumption in the process of obtaining the time. For instance, the computer CPU often had to be tied up with timing events when other work required being done in the timed interval.

A wonderful solution to the timing problem eventually presented itself in what the computer people call a "real-time-clock" which can be added to the "IBM-type" microcomputers for as little as $50. Operating on a small battery it maintains the correct date and time even when the computer is turned off. With appropriate software, the "real-time-clock" device allows the timing of events accurately to the nearest 1/1000-th of a second with negligible loss of computer time in the reading. (The sub-program used in our projects will read the time in approximately 1/3000-th of a second.)

<u>Peripheral Devices for Response Acquisition: Response Pedestal.</u> The initial choices in the hardware configuration for a "testing station" proved satisfactory for the "stimulus side", i.e., the controlled presentation to the subject. The standard keyboard and the "off-the-shelf" joysticks were hopelessly inadequate for the "response side." Computer keyboards leave much to be desired as response acquisition devices--particularly when response latency is a variable of interest. Preliminary trials using, say, the "D" and "L" keys of the keyboard for "true" and "false" responses to items was troublesome with naive subjects. Intricate training was required to avoid individual differences arising from differential experience with keyboards. Moreover, the software had to be contrived so as to flash a warning when a respondent accidentally pressed any other key. The "off-the-shelf" joysticks were sadly lacking in precision of construction such that the score of a respondent depended heavily upon which joystick she/he was using.

We came up with a plan for a "response pedestal" which consisted of readily available electronic parts. A prototype of the device was obtained from a local engineer. (See Figure 1.) It had two joysticks, a horizontal and a vertical sliding adjuster, and a dial. The two joysticks allowed either left or right hand usage. The sliding adjusters permitted two-handed coordin-

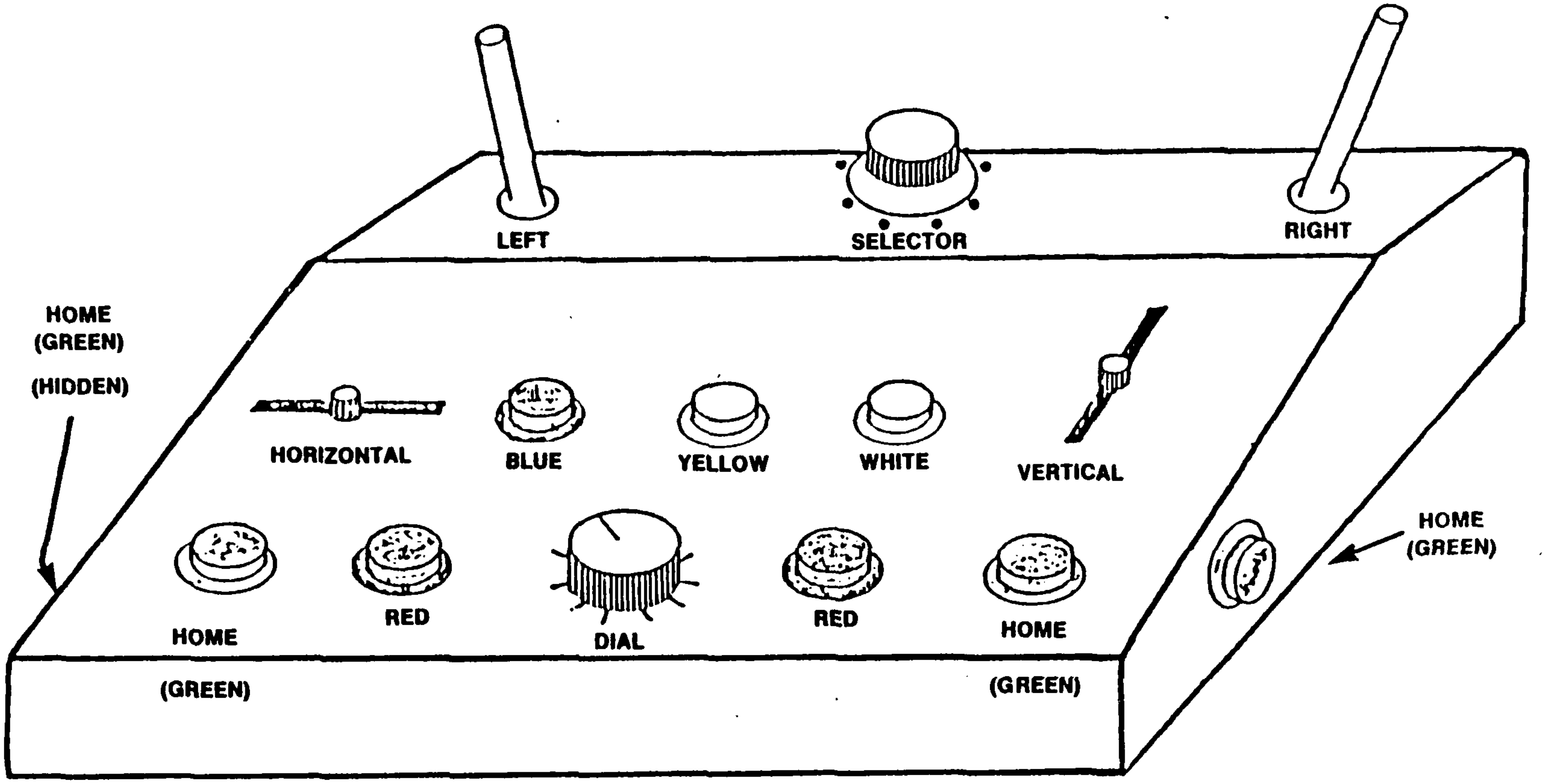


FIGURE 1. Custom-designed response pedestal

ation tasks. The dial permitted respondent selections in a manner similar to the now popular "mouse" devices that are sold for "personal computers."

The response pedestal had nine button-switches, each of which was to be used for a particular purpose. Three buttons (BLUE, YELLOW, and WHITE) were located near the center of the pedestal and were used for registering up to 3-choice alternatives. Also near the center were two buttons (RED) which were mostly used to allow the respondent to step through frames of instructions and, for some tests, to "fire" a "weapon" represented in graphics on the screen.

Of notable interest was the placement of the button-switches which were called "HOME" with respect to the positions of other buttons used to register a differential response. The "HOME" buttons required the respondent's hands to be in the position of depressing all four of the "HOME" buttons prior to presentation of an item to which (s)he would respond. This, it is believed, offered advantages of control of attention and control of hand position for measurement of response latency. Using appropriately developed software, we were able to measure total response time but also to break it down into two parts: (1) "decision time" which is defined as the interval between onset of stimulus and release of the "HOME" keys, and (2) "movement" time which is the subsequent interval to the registering of a response. It was possible, where of interest, to even tell quite reliably whether the respondent used a left hand or a right hand to respond since (s)he almost invariably would release the "HOME" buttons on the side of the preferred hand first.

The rotary switch marked "SELECTOR" in Figure 1 was an inconvenience that was required by our initial choice of "game-adapter" for reading analog input. The game adapter initially chosen allowed only four inputs and the response pedestal had seven analog outputs: 2 inputs for each of two joysticks, two sliding adjusters, and one rotary adjuster called the "DIAL." The "SELECTOR" was used to select which analog devices were to be operative for a particular test item. The final design for the response pedestal included a game-adapter with the capability of eight analog inputs and the "SELECTOR" switch was happily omitted.

Joysticks. Perhaps the greatest difficulty regarding the response pedestal design arose from the initial choice of joystick mechanisms. We soon discovered that joystick design is a complicated and, in this case, a somewhat controversial issue. Variations in tension or movement can defeat the goal of standardized testing. While "high-fidelity" joystick devices are available, they can cost thousands of dollars apiece which was prohibitively expensive in the quantities that were to be required for this project. The first joystick mechanism that was used in the response pedestals was an improvement over the initial "off-the-shelf" toys that predated the pedestals. It had no springs whatsoever so that spring tension would not be an issue. It had a small, light weight handle so that enthusiastic respondents could not gain sufficient leverage to break the mechanism. It was inexpensive.

Unfortunately, this joystick had a "wimpy" feeling which was greatly lacking in "face-validity" (or, as Hilda Wing dubbed it, "fist-validity") from the Army's point of view. It was felt that the joystick was so much like a toy that it would not command respect of the respondents. It was the contention of a minority of us that our "wimpy" device had "construct fidelity" in that it would do a perfectly adequate job of testing the constructs that were targeted.

The joystick mechanism had to be changed. Joysticks of every conceivable variety and type of use were considered. We learned about viscous dampening, friction, tension, and even something called "stiction." Ultimately, a joystick device was fashioned with a light spring for centering and a sturdy handle with a bicycle handle-grip. It had sufficient "fist-validity" to be accepted by all (or almost all) and it was sufficiently precise in design that we were unable to detect any appreciable "machine" effects in fairly extensive testing.

Software Development

We wish to turn attention now to the issues of software development. There were no "package programs" available to administer computerized tests. The selection of strategy for organizing and programming the needed software was to fall upon ourselves. We had three general, operational objectives in mind for the software to be produced: (1) as far as possible, it should be transportable to other microprocessors; (2) it should require as little intervention as possible from a test administrator in the process of presenting the tests to subjects and storing the data; and, (3) it should enhance the "standardization" of testing by adjusting for hardware differences across computers and response pedestals.

Primary Language. We chose to prepare the bulk of the software using the Pascal language as implemented by Microsoft, Inc. There were certain advantages to this in that Pascal is a common language and it is implemented using a compiler that permits modularized development and software libraries. As computer languages go, Pascal is relatively easy for others to read and it can be implemented on a variety of computers.

Some processes, mostly those which are specific to the hardware configuration had to be written in IBM-PC assembly language. Examples of these include the interpretation of the response pedestal inputs, reading of the real-time-clock registers, calibrated timing loops, and specialized graphics and screen manipulation routines. For each of these identified functions, a Pascal-callable "primitive" routine with a unitary purpose was written in assembly language. The functions were designed to be simple and unitary in purpose so as to be easily reproducible for other machines.

Strategy. The overall strategy of the software development is worth discussing. It quickly became clear that the direct programming of every item in every test by one person was not going to be very successful either in terms of time constraints nor in terms of quality of product. For the sake of making it possible for each researcher to contribute his/her judgment and effort to the project, it was necessary to plan so as to take the "programmer" out of the step between conception and product as much as possible.

The testing software modules were designed as "command processors" which interpreted relatively simple, problem oriented commands. These were organized in ordinary text written by the various researchers using word processors. Many of the commands were common across all tests. For instance, there were commands that permitted writing of specified text to "windows" on the screen and controlling the screen attributes (brightness, background shade, etc). A command could hold a display on the screen for a period of time (measured to 1/100-th second accuracy). There were commands which caused the program to wait for the respondent to push a particular button on the pedestal. Some of the commands were specific to particular item types. These commands were selected and programmed according to the needs of a particular test type. For each item type, we would decide upon the relevant stimulus properties to vary and build a command that would allow the item writer to quickly construct a set of commands for items which she/he could then inspect on the screen.

Thus, entire tests were constructed and experimentally manipulated by psychologists who could not program a computer.

The strategies for developing commands have evolved and improved over the period of development. Eventually, the commands became almost "language-like" with syntax forms analogous to some of the common statistical packages like SPSS or SAS that are available on "main-frame" computers.

Hardware Testing and Calibration. One of the most useful software developments relates to the testing and calibration of the hardware, necessary for purposes of standardization. A complete hardware testing and calibration process can be undertaken by test monitors each time a machine is powered up. It checks the timing devices and screen distortion, and calibrates the analog devices (joysticks, sliding adjusters, dial) so that measurement of movement will be the same across machines. It also permits the software adjustment of the height to width ratio of the screen display so that circles do not become ovals or, more importantly, the relative speed of moving displays remains under control regardless of vertical or horizontal travel.

Concluding Remarks

In the end, we were able to put together a portable, complete testing session lasting approximately 1-1/2 hours where very naive respondents can complete the test with little or no intervention from a test monitor. The data is automatically stored and "backed-up" on diskettes in a form readily transferrable to a "main-frame" for analysis. Except for occasional calibration or contingencies, the test monitor needs only to turn the computers on and put the respondents in front of them.

Finally, and perhaps most gratifying, we have found that the soldiers tested via this method have generally preferred computerized testing to paper-and-pencil testing. We have not gathered hard data on this aspect, but base our conclusions on observation of the soldiers while taking the battery and their comments to us after completing the battery. Perhaps this is due to novelty alone, but we feel it may also be due to the nature of the tests themselves plus the fact that the soldier, in large part, is in control of the testing process her/himself. They control the pacing of instructions for the tests and, for some tests, the pacing of item presentation. No administrator tells them when to begin and when to stop, and they are not in "lock step" with a larger group. We view this state of affairs as highly desirable for personnel selection testing.

THE VALIDITY OF ASVAB FOR PREDICTING TRAINING AND SQT PERFORMANCE

Paul G. Rossmeissl
U.S. Army Research Institute

Donald H. McLaughlin, Lauress L. Wise, and David A. Brandt
American Institutes for Research

Presented on symposium,
"Entry Standards Based on the ASVAB"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

The Validity of ASVAB for Predicting Training and SQT Performance

Paul G. Rossmeissl
U.S. Army Research Institute[1]

Donald H. McLaughlin, Lauress L. Wise and David A.Brandt
American Institutes for Research

This paper is a condensation of a larger report (McLaughlin, Rossmeissl, Wise, Brandt, & Wang; 1984) which investigated the validity of the Armed Services Vocational Aptitude Battery (ASVAB) for predicting success in Army jobs or Military Occupational Specialties (MOS). The ASVAB is a cognitive test battery used by the military services as their primary instrument for selecting and classifying enlisted personnel. This particular research was based upon ASVAB forms 8/9/10 which was composed of ten subtests: General Sciences (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto/Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehensive (MC), and Electronics Information (EI). The two verbal subtests, WK and PC, are most often combined into a single measure of verbal ability called VE. The current version of ASVAB (forms 11/12/13) uses parallel forms of these same subtests.

Scores on the ten ASVAB subtests are typically combined into aptitude area (AA) composites. Examples of these composites are given in Table 1. The Army composites serve as the basis for assignment of personnel to Army MOS in that a minimum qualifying score on one of the aptitude area composites is required for admission to Army initial level training courses. For example, the CO composite is used to classify recruits into the infantry and armor specialties. Similarly, the MAGE composites are used by the Air Force to select and classify prospective personnel into Air Force specialties. The final set of composites routinely in use are the High School Composites which have been developed for use when ASVAB is administered to high school students as a career guidance tool. Maier and Truss (1983) have also recommended that the first four of these composites be used to select and classify enlisted personnel within the Marine Corps.

The goal of the McLaughlin et al. (1984) research was twofold. First, the validities of the composites then in use by the Army and other DoD agencies were evaluated with regard to predicting success within the Army. Second, an additional set of composites were derived empirically in hopes of obtaining a composite system with maximal predictive validity.

In all cases the validation criterion were MOS specific end-of-course training scores or skill qualification tests (SQTs). All of the criterion measures were trimmed of outliers and then standardized before any

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

validation analyses. The separate training and SQT data were combined for validation analyses at the MOS level. All validities were corrected for restriction of range using the multivariate adjustment due to Lawley (1943) and described by Lord and Novick (1968).

Table 1
Typical ASVAB Composites

Army Composites (1983)		
Clerical/Administrative	CL	VE + NO + CS
Combat	CO	AR + CS + AS + MC
Electronics Repair	EL	GS + AR + MK + EI
Field Artillery	FA	AR + CS + MK + MC
General Maintenance	GM	GS + AS + MK + EI
Mechanical Maintenance	MM	NO + AS + MC + EI
Operators/Food	OF	VE + NO + AS + MC
Surveillance/Communications	SC	VE + NO + AS + CS
Skilled Technical	ST	VE + GS + MK + MC
MAGE Composites		
Mechanical	M	MC + AS + GS
Administrative	A	VE + NO + CS
General	G	AR + VE
Electronic	E	AR + MK + GS + EI
High School Composites		
Mechanical Trades	HSMT	AR + MC + AS + EI
Office and Supply	HSOS	VE + CS + MK
Electronics/Electrical	HSEE	AR + EI + MK + GS
Skilled Services	HSSS	AR + VE + MC
Academic Ability	HSAA	AR + VE

Composite System Validities

Table 2 gives the adjusted validities for each of the composite systems displayed in Table 1. Validities and sample sizes are given for each of the nine clusters of MOS now in use by the Army. The validities were obtained by averaging the validities for the individual MOS within each cluster and weighting by the number of soldiers within each MOS.

Table 2
Validities of Established Composite Systems

Army Composites (1983)

Cluster					Composite					
of MOS	(N)	CL	CO	EL	FA	GM	MM	OF	SC	ST
CL	10368	48	51	53	54	49	46	50	50	53
CO	14266	36	44	43	43	43	42	44	40	44
EL	5533	38	47	46	47	46	47	44	44	47
FA	5602	39	49	48	48	49	49	49	45	44
GM	2571	39	48	46	46	47	48	48	45	47
MM	7073	36	48	46	45	48	48	48	43	46
OF	8704	38	48	47	45	48	47	48	44	48
SC	3729	39	49	48	47	48	47	48	45	49
ST	7061	51	56	57	57	55	54	56	54	58

MAGE Composites

Cluster			Composite		
of MOS	(N)	M	A	G	E
CL	10368	45	48	54	53
CO	14266	42	36	42	43
EL	5533	45	38	46	47
FA	5602	48	39	46	48
GM	2571	46	39	44	46
MM	7073	48	36	44	46
OF	8704	47	38	47	47
SC	3729	47	39	47	48
ST	7061	52	51	57	57

High School Composites

Cluster				Composite		
of MOS	(N)	HSAA	HSMT	HSOS	HSSS	HSEE
CL	10368	54	47	54	53	53
CO	14266	42	43	40	44	43
EL	5533	46	47	43	47	47
FA	5602	46	49	44	49	48
GM	2571	44	47	43	47	46
MM	7073	44	49	41	47	46
OF	8704	47	48	43	48	47
SC	3729	47	48	44	49	48
ST	7061	57	54	56	58	57

The main diagonal of the upper portion of Table 2 gives the validities of the composites that were associated with each of the nine clusters in 1983. The most interesting feature of the data in Table 2 is the uniformity of the validities. All of the entries are between .36 and .58, with the mean validity of each system being about .45. One MOS cluster, ST, appears to be slightly more predictable than the others; and another cluster, CO, appears to be slightly less predictable. The remaining clusters show very little variance.

Identification and Validation of Alternative Composites

In order to develop alternative composites the MOS were partitioned into clusters, based on similarity of ASVAB profiles of successful criterion performance. The similarity between each pair of cells was defined as correlation of the predicted criterion performance in the two cells for the applicant sample. The performance predictions were based on ridge regressions, using the ASVAB subtests as predictors. The cells were clustered by adapting standard "leaf to stem" procedures. Upon finding that the results of the clustering were unstable, due to the high intercorrelations of the predicted criterion scores, the clustering procedure was modified to use as a starting point the Army's current grouping of MOS into aptitude area clusters.

Once a cluster had been defined the unit-weight composite with maximal predictive validity for that cluster was identified. It was found that optimal unit-weight composites for four clusters possessed a root mean square (RMS) predictive validity within 97% of the RMS validity of the ridge regression vectors computed separately for each of the 98 MOS included in the sample. The composition of these four alternative composites are given in Table 3, and their predictive validities are given in Table 4.

Table 3
Optimal Four Composite Solution

Composite		Subtests
Clerical/Administrative	(ACL)	VE + AR + MK
Skilled Technical	(AST)	VE + AR + MK + AS
Operations	(AOP)	VE + AR + MC + AS
Combat	(ACO)	VE + MK + MC + AS

Inspection of Table 4 shows that by focusing on the most valid portion of the ASVAB, the primary aim of this aspect of the research was achieved: the validities went up. The aggregate RMS predictive validity for the four alternative composites for their assigned MOS is .486, in comparison with RMS validity for the 1983 Army composites of .454. Certain members of the 1983 Army composite set account for a large part of the difference in validity between the two composite sets. When compared

Table 4
Predictive Validities of the Alternative Composites

Cluster of MOS	(N)	ACL	Composite AST	ACO	AOP
CL/ACL	10368	56	54	52	51
CO/ACO	14266	42	44	44	44
EL/ACO	5533	46	48	48	48
FA/ACO	5602	47	49	50	50
GM/ACO	2571	45	48	48	48
MM/AOP	7073	44	48	49	49
OF/AOP	8704	46	49	49	49
SC/AOP	3729	47	49	50	50
ST/AST	7061	58	58	57	57

to validities of the optimal composites for the same cluster of MOS, the 1983 Clerical composite (CL) appeared to be weak, with a validity of .48 versus a potential of .56. Another composite Surveillance and Communications (SC), was mildly weak, with a validity of .45 versus a potential .50.

Recommendations

A major purpose behind the McLaughlin et al. (1984) report was to present recommendations to the Army as to how the composite system then in use to select and classify enlisted personnel could be improved. The average validity of the set of four empirically derived alternative composites was .48 versus .45 for the existing composite systems. Thus, from a purely statistical point of view the results in terms of predictive validity tended to favor the alternative four composite solution over the nine composite system the being used or any of the alternatives being used by other armed services.

However, considering the costs of implementing a whole new composite system, it was decided that a more favorable proposal would be to maintain a nine composite system but to replace the the two composites which were the major source of the deficiency of the 1983 composites. The new CL composite would be comprised of the VE, AR, and MK subtests and would have a predictive validity of .56. The new SC composite would have a predictive vaidity of .50 and be made up of the VE, AR, MC, and AS subtests. The average validity of the revised nine composite system would be .47. The Army officially adopted this composite system on October 1, 1984.

The gain in expected performance resulting from the change in the CL and SC composites can only be approximated, because of the constrained nature of the selection and classification process. If, however, the choice were purely between assignment to an individual MOS and rejection, application of Cronbach's formula yields an expected gain of .05 standard deviations of criterion performance per person in the two clusters of MOS from the introduction of the two revised composites.

References

Lawley, D. (1943). A note on Karl Pearson's selection formulae. Royal Society of Edinurgh, Proceedings, Section A. 62, 28-30.

Lord, P., & Novick, M. (1968). Statistical theory of mental test scores. Reading MA: Addison-Wesley Publishing Company, Inc.

Maier, M. H., & Truss, A. R. (1983). Validity of ASVAB Forms 8, 9, and 10 for Marine Corps training courses: Subtests and current composites. (Center for Naval Analyses Memorandum No. 83-3107). Alexandria, VA: Marine Corps Operations Analysis Group.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A. & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY 1981 and FY 1982 enlisted accessions. Technical Report No. 651, U. S. Army Research Institute for the Behavioral and Social Sciences, Alexandria VA.

DEVELOPING NEW ATTRIBUTE REQUIREMENTS SCALES FOR MILITARY JOBS

Elizabeth P. Smith
U.S. Army Research Institute

Presented on symposium,
"Determining Ability Requirements"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Developine New Attribute Requirements Scales for Military Jobs

Elizabeth P. Smith
U.S. Army Research Institute[1]
5001 Eisenhower Avenue
Alexandria, Virginia 22333-5600

Conducting empirical validity investigations to predict job performance is not always feasible. Even when empirical approaches are undertaken, such as the ongoing ARI Project A to improve the selection, classification and utilization of enlisted personnel, it is rarely possible to include all jobs within an organization. Given the complexities of empirical validation, it is necessary to develop other methods for matching people to jobs and optimizing their performance.

One approach is to obtain rational estimates of the human attributes (i.e., abilities, characteristics, and interests) which are required for successful job performance. When gathered systematically from qualified judges, these estimates can be summarized as profiles of required attributes. Then, measures of individuals' attributes can be matched to such profiles for selection and classification purposes. In addition, knowledge of required attributes is potentially useful for (a) designing new systems and training programs that are within the capacities of available personnel and (b) generalizing empirical validity data to new and different jobs, by grouping them on the basis of similarity of attribute profiles (Fleishman, 1982; Pearlman, 1980). The latter application is especially pertinent to the Army's Project A, which is collecting validity data for only 19 Military Occupational Specialties (MOS).

A well-researched method of determining ability requirements is the rating scale approach developed by Fleishman and his associates (see Fleishman & Quaintance, 1984 for a comprehensive summary), based on a taxonomy of 40 cognitive, perceptual, physical and psychomotor abilities. With these scales, a rater decides if an ability is necessary for errorless job performance, and, if so, estimates the level required on a 7-point, behaviorally-anchored scale.

Early outcomes from Project A provided an opportunity to develop a new set of rating scales based on a new taxonomy of human attributes. An expert judgment task (Wing, Peterson, & Hoffman, 1984) obtained estimates of validity for 53 predictors against 72 criterion constructs from 35 personnel psychologists. Factor analysis of the data yielded 21 clusters of the 53 cognitive, perceptual, psychomotor, temperament and interest predictor variables. A predictor test battery based on these 21 clusters has been developed and is being validated. The purpose of this paper is to discuss the initial construction and testing of a new set of scales for estimating job requirements which is based on these 23 clusters (hereafter called "attributes"). As more data become available, it is expected that the taxonomy of predictors (and test battery) may change. The rating scales will be revised to reflect these changes.

A set of scales based on the Project A taxonomy has several potential advantages over the Fleishman ones. The most salient feature is that obtained profiles of attribute requirements will directly correspond to

[1] The views expressed in this paper are those of the author and do not necessarily reflect the view of the U. S. Army Research Institute or the Department of the Army.

P: ject A validity data. It will include temperament and interest measures that are not among the Fleishman scales and will not include those attributes/abilities for which no predictor tests are given. Additional benefits (e.g., lower cost, more efficiency) may be possible with this set of scales. It was designed to be used by work supervisors rather than personnel psychologists and contains primarily Army-specific behavioral anchors with only about half as many attributes to rate as Fleishman's.

For any rating scales to be useful in practice, they must give reliable and valid scores. The effort reported here examined issues related to the reliabilitv of the ratings. Validity investigations will occur later. The following issues were examined here. First, how closely do raters agree, i.e., how high is interrater reliability? Second, how well do the scales differentiate across attributes (i.e., yield non-flat profiles) within a job and across the attribute profiles of different jobs? Finally, can the scales be used to identify attributes for which differences in level of the attribute most influence performance? For some attributes, higher levels may be required for better performance whereas for others, once a minimal requirement is met, having a greater amount of the attribute has no additional effect on performance.

METHOD

Subjects. Thirty-six Non-commissioned Officers (NCOs) from the Cannon Crewman MOS and 39 NCOs from the Motor Transport Operator MOS, all males located overseas, participated as Subject Matter Experts (SMEs).

Instrument. The Attribute Assessment Scale, which was empirically developed for this research, consists of a set of behaviorally-anchored scales for 20 of the 21 attributes in the Project A taxonomy plus two additional attributes, Stamina and Physical Strength, which were thought to enhance face validity. A scale for Enterprising Interests was eliminated because it was impossible to generate items for this attribute which were sufficiently different from those falling under Self-Esteem/Leadership. The names of the attributes were modified from the original Wing, et. al. (1984) labeling for better comprehension by SMEs. The final instrument had one page per attribute. Below the definition at the top, there were three 7-point vertical scales, placed side-by-side, to enable three responses. A zero-point was added to indicate the attribute was not required at all. SMEs circled the number corresponding to the appropriate level for their job.

To construct the scales, comprehensive definitions for the attributes were developed so as to be readily understandable by people who were not trained in personnel research. A pool of items for potential anchors (i.e., behavioral statements) was generated. Ten items per attribute were ultimately selected, after screening by two to four other researchers. These were presented with the appropriate definition in an anchor-rating instrument. Initially, 26 NCOs from either the Administrative Specialist or Military Police MOS rated each item on the amount of the attribute represented by or needed for the behavior described. Items with mean ratings that were the highest, lowest, and closest to 4.0 (midpoint) that also had a standard deviation less than 1.5 were selected as scale anchors. Using these criteria, scales could be created for only 11 attributes.

After identifying difficulties related to (a) task comprehension, (b) response format, (c) failure of raters to differentiate effectively among items, and (d) a few of the definitions and items themselves, I revised the anchor-rating instrument and administration procedures, adding a

15-minute training period. This instrument was given to another sample of NCOs (*N*=28) from the same two MOS. From the second administration, using the criteria indicated above, three anchors were obtained for all but two of the attributes (Social Interaction and Stress Reaction for which only two anchors were selected) to form the Attribute Assessment Scale.

Procedure. SMEs rated the level of each of the 22 attributes that is required to perform Skill Level 1 (entry level) work under combat-readiness conditions in his own MOS for three performance levels: at the 15th, 50th, and 85th percentiles. In addition to the written instructions, SMEs received extensive training in how to complete the task, including a step-by-step demonstration of the actual rating process using the anchors as guides. Training and responses to questions took about an hour. Early ratings were checked to ensure comprehension of the directions before raters proceeded with the rest of the task. Ratings took about 30-45 minutes.

Analyses. Intraclass correlation coefficients (ICCs) were calculated from Raters X Attributes ANOVAs over all attributes and separately for the three major domains (i.e., cognitive/perceptual, physical/psychomotor, and noncognitive) for each of the three performance levels. The ICCs estimate the reliability of the mean ratings [r(k); k=number of raters], an index of interrater reliability. Also, an MOS X Attributes X Performance Levels univariate repeated-measures ANOVA was performed.

RESULTS

Eight Motor Transport Operators were eliminated from the analyses due to the logical inconsistency of their data. *R*(k) coefficients over all attributes were, in increasing order by performance level, .75, .77, and .69 for Cannon Crewmen (k=36) and .74, .74, and .69 for Motor Transport Operators (k=31). For the domains, r(k) coefficients ranged from .61 to .79 across performance levels and MOS. There were two exceptions to this: Physical/psychomotor reliabilities were very low for both MOS at the 85th percentile [*r*(36)=.13; *r*(31)=.38] performance level.

None of the effects involving MOS for the MOS X Attributes X Performance Levels ANOVA were significant. There were significant main effects for Attributes [*F* (21,1365) = 6.98; *p* = .0000] and Performance Levels [*F* (2,130) = 398.36; *p* = .0000] and a significant effect for the Attributes X Performance Levels interaction [*F* (42,2730) = 2.51; *p* = .0000]. Scheffe' comparisons between means within performance levels by MOS indicated significant differences between only the highest and lowest means, which ranged from 1.09 to 1.75. Means and standard deviations for all ratings are provided in Table 1.

DISCUSSION

In comparison to the very high Intraclass Correlation Coefficients (ICCs) obtained by Fleishman and associates or those discussed by Rossmeissl (1985) within this symposium, the ICCs from this research are weak, especially since around 30 raters are needed to obtain coefficients of at least .60. ICCs are based on variance components. As such, low (or uninterpretable) reliabilities result if there is too great a between-subjects variance and/or too little within-subjects variance. The low reliabilities obtained here appear to be a function of both. Previous research on ability assessment has found mean ratings that varied from very low (even "Not required") to very high (7) across attributes. This was not the case here. The inclusion of three performance levels may have

Table 1

Mean and Standard Deviations of Ratings of Attribute Requirements for Cannon Crewman and Motor Vehicle Operator MOS at Three Performance Levels.

Attributes	MOS[a]	Performance Level					
		15th Percentile		50th Percentile		85th Percentile	
Cognitive		M	SD	M	SD	M	SD
Verbal Ability	C	2.86	(.93)	4.17	(.91)	5.33	(1.10)
	D	2.87	(.83)	4.32	(.85)	5.32	(1.08)
Memory	C	2.44	(1.18)	3.89	(.89)	5.53	(1.11)
	D	3.26	(1.26)	4.26	(1.03)	5.16	(1.27)
Reasoning Ability	C	2.78	(1.31)	3.94	(1.17)	4.89	(1.33)
	D	2.45	(.96)	3.77	(1.02)	5.00	(1.32)
Number Facility	C	2.06	(1.12)	3.47	(1.16)	5.08	(1.52)
	D	2.48	(1.18)	3.90	(1.30)	4.94	(1.48)
Mechanical Comprehension	C	2.86	(1.36)	4.39	(1.18)	5.50	(1.08)
	D	2.97	(1.33)	4.23	(1.14)	5.20	(1.00)
Information Processing	C	2.50	(1.08)	3.58	(1.30)	4.81	(1.37)
	D	2.68	(1.30)	3.90	(1.08)	5.03	(1.17)
Closure	C	2.78	(1.33)	4.08	(1.25)	4.89	(1.41)
	D	2.68	(1.42)	3.87	(1.45)	4.61	(1.67)
Visualization	C	2.33	(1.15)	3.58	(1.30)	4.69	(1.51)
	D	2.29	(1.30)	3.42	(1.43)	4.26	(1.84)
Perceptual Speed & Accuracy	C	2.75	(1.52)	3.97	(1.38)	4.94	(1.31)
	D	2.97	(1.33)	4.16	(1.29)	4.71	(1.40)
Physical/Psychomotor							
Physical Strength	C	3.75	(1.32)	4.89	(1.14)	5.67	(1.17)
	D	3.58	(1.39)	4.61	(1.20)	5.32	(1.25)
Stamina	C	3.06	(1.45)	4.53	(1.11)	5.69	(1.19)
	D	2.68	(1.30)	3.94	(1.41)	4.84	(1.63)
Multilimb Coordination	C	2.80	(1.47)	4.20	(1.21)	5.26	(1.40)
	D	3.34	(1.54)	4.45	(1.24)	5.48	(1.12)
Dextery	C	3.00	(1.35)	4.47	(1.08)	5.50	(1.08)
Non Cognitive							
Steadiness/Precision	C	2.83	(1.40)	4.08	(1.25)	5.47	(1.36)
	D	3.06	(1.29)	4.52	(1.09)	5.29	(1.07)
Social Interaction	C	3.14	(1.62)	4.44	(1.59)	5.34	(1.70)
	D	2.58	(1.71)	3.74	(1.44)	4.65	(1.70)
Stress Tolerance	C	3.03	(1.50)	4.22	(1.27)	5.12	(1.43)
	D	3.10	(1.47)	4.27	(1.34)	5.27	(1.20)
Conscientiousness	C	2.66	(1.24)	4.09	(.89)	5.31	(.99)
	D	3.19	(1.47)	4.35	(1.11)	4.97	(1.25)
Work Orientation	C	2.91	(1.46)	4.29	(1.18)	5.54	(1.17)
	D	2.90	(1.45)	4.16	(1.10)	5.39	(1.17)
Self Esteem/Leadership	C	3.00	(1.26)	4.25	(1.20)	5.47	(1.21)
	D	2.48	(1.55)	3.84	(1.37)	5.00	(1.41)
Athletic Ability/Energy	C	2.89	(1.35)	3.92	(1.16)	4.94	(1.19)
	D	2.87	(1.55)	3.73	(1.48)	4.33	(1.71)
Realistic Interests	C	2.54	(1.40)	3.71	(1.25)	4.94	(1.66)
	D	2.48	(1.12)	3.61	(.88)	4.61	(.99)
Investigative Interests	C	2.00	(1.43)	3.44	(1.61)	4.61	(1.78)
	D	1.97	(1.40)	3.26	(1.50)	4.16	(1.93)

[a] C = Cannon Crewman

D = Motor Vehicle Operator (Driver)

had a strong, negative impact on these particular results. The demands of the task appeared to impose a unique kind of restriction in the range of possible ratings. That is, the effective range of ratings within levels covered only two or three points rather than the entire seven points. This outcome served to reduce within-subjects variability, as all ratings fell close together. Although SMEs were clearly advised not to respond according to belief that "better must mean more," the mean ratings suggest that a demand characteristic was created by the instructions to rate at three levels. The result was ratings of attribute levels which correspond to level of performance, with ceiling effects occurring at the highest level. These effects would explain the extremely low reliabilities for Physical/Psychomotor attributes at the 85th percentile.

The fact that attribute requirements were elicited for three performance levels also may have clouded the findings in another way and reduced interrater agreement, i.e., increased between-subjects variance. Although definitions were provided for the three performance levels, how the SMEs actually interpreted these definitions was unknown. SMEs may have had different interpretations of the attributes from our definitions as well as from one another. For example, their verbal reports seemed to indicate some tendency to interpret performance levels in terms of particular soldiers in their charge, rather than from a more general (and shared) view of job performance at a particular level. They also tended to rate attributes in terms of the characteristics of someone who performed at that level, rather than in terms of the actual requirements of the job. The performance criterion, then, was more ambiguous than expected, pointing out a clear need for a very specific definition of the criterion. It was apparent that understanding the task requirements -- what was meant by the performance levels and how to do three ratings at a time -- took more time and energy than actually doing the ratings. In short, the use of three performance levels may have made the task harder than was intended, and interfered with the SMEs' ability to rate true requirements.

Two other factors may have contributed to low interrater agreement. SMEs were not given written descriptions of what they were to rate. Instead they were asked to decide individually the nature and content of entry level work and, specifically, what it required in terms of attributes. Moreover, they were to rate the whole job -- all work within all duty positions -- and not just some specific task or set of tasks. This very broad scope allowed considerable opportunity for variance. As a result of personal experiences and/or selective memory, the SMEs could differ a great deal in what they were evaluating. Obviously, higher interrater agreement would be expected for narrower areas of consideration. In addition, some SMEs found the scale anchors frustrating rather than helpful. Raters appeared to have difficulty using anchors as reference points for comparing tasks within their MOS. Some tended to evaluate the job in terms of whether the exact tasks depicted were or were not an actual part of the job. With some anchors that depicted common soldier tasks, some SMEs had problems separating the overall soldier requirements from the specific job requirements. Thus, although very familiar behaviors were thought to be the best for illustrating a level of an attribute, this was not necessarily the case.

The results of the ANOVA indicate that attribute profiles for the two MOS are not significantly different. The effects that were significant, Attributes, Performance Levels, and their interaction, are most likely a function of the high statistical power related to the large number of degree of freedom, and so are not really meaningful. Despite this, the data provide some useful information. The minimal differences which do occur suggest that some differences (as well as similarities) between MOS may exist, but may be masked in the present research. In addition, rank orders of the magnitude of ratings were different for both MOS at all performance levels, again suggesting there may be some differences in patterns of attributes which need further examination. For instance, at the 85th percentile, Verbal Ability ranked tenth for Cannon Crewman but third for Motor Vehicle Operator, while Stamina ranked first and fifteenth respectively. If one were to select only the five variables with the highest ratings, the selection would be different for each MOS. However, the top five are not necessarily the most important attributes: They are ranked on level of required attribute and not on relative importance of the attribute.

In summary, NCOs appeared to understand, in general, how to use the set of scales constructed to rate job requirements. The requirement for three sets of ratings simultaneously, however, created some problems. First, the actual physical arrangement of the scales on the page confused people. Second, it seemed to impose limits on the magnitude of ratings assigned. Given the expanse of the criterion to be rated -- the entire MOS at Skill Level 1 -- and the limitations created by the design itself -- different performance levels -- the obtained indices of interrater agreement are reasonable.

These findings suggest that better reliability estimates could be obtained with fewer raters if SMEs were asked to rate requirements for a single performance level; i.e., to estimate the minimum level of an attribute required to perform the job successfully. Further,' more reliable ratings may be obtained by changing to a generic set of scale anchors (e.g., very low,low, moderate, etc.) or otherwise replacing the present behavioral anchors and/or focusing raters' attention on evaluating a specific task, a well-defined set of tasks, or a written job description would yield better reliability coefficients. Elimination of the restriction in range of ratings which was created by including three performance levels, should yield better discrimination among the attributes within MOS, and differences in attribute profiles across MOS.

References

Fleishman, E. A. (1982). Systems for describing human tasks. American Psychologist. 30, 1127-1149.

Fleishman, E. A., & Quaintance, M. K (1984). Taxonomies of human performance. Orlando, FL: Academic Press Inc.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87, 1-28.

Rossmeissl, P. G. (1985, October). Computerized approaches for estimating ability requirements. Paper presented at the 26th Annual Conference of the Military Testing Association, San Diego, CA.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1984, August). Expert judgments of predictor-criterion validity relationships. Paper presented at the 92nd Annual Convention of the American Psychological Association, Toronto, Canada.

ADDING TO THE ASVAB:
COGNITIVE PAPER-AND-PENCIL MEASURES

Jody L. Toquam, Marvin D. Dunnette, VyVy A. Corpe, and
Janis Houston

Personnel Decisions Research Institute

Presented on symposium,
"Predicting a Broad Variety of Criteria:
Elaborating the Predictor Space"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

ADDING TO THE ASVAB: COGNITIVE PAPER-AND-PENCIL MEASURES

Jody L. Toquam, Marvin D. Dunnette, VyVy A. Corpe, and
Janis Houston

Personnel Decisions Research Institute

Introduction

The purpose of this paper is to (1) identify the cognitive/perceptual ability constructs that supplement or provide information about Army applicants' abilities not currently tapped by the Armed Services Vocational Aptitude Battery, or ASVAB; (2) describe the measures developed for paper-and-pencil administration and the cognitive/perceptual constructs they are designed to tap: (3) describe test development issues and the factors used to evaluate the psychometric quality of the new paper-and-pencil measures; and (4) report the relationships between scores on the ASVAB and scores on the new paper-and-pencil tests. Information about the cognitive/perceptual measures designed for computer administration are described in McHenry and Toquam (1985).

Before describing the new tests, we first examine the content of the current military selection and classification battery, the ASVAB, and then provide a brief review of the process involved in identifying the constructs for inclusion in the Pilot Trial Battery. (The Pilot Trial Battery is the term used for the battery of experimental tests administered at Minneapolis MEPS, Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. This battery includes twelve paper-and-pencil measures - ten cognitive and two non-cognitive, and ten computerized measures - seven cognitive/perceptual and three psychomotor.)

The current military selection and classification battery, the ASVAB, contains ten subtests. Scores on four of these are used to calculate the Armed Forces Qualification Test (AFQT) score which is used to determine qualification for entrance into the Army. Scores on the ten subtests are used in different combinations to determine applicants' qualifications for different military occupational specialties (MOS). Results from a factor analysis of ASVAB scores indicate that the battery assessed verbal ability, speeded performance, quantitative ability, and technical knowledge (Kass, Mitchell, Grafton & Wing, 1982).

Peterson (1985) describes the activities involved in identifying ability constructs that supplement information obtained from the ASVAB. Those activities included a review of the literature which was used to impose structure on the domain (i.e., establish a cognitive/perceptual abilities taxonomy) and then to summarize validity data for the different types of ability constructs. This information was input to the expert judgment task. All of this information was used to identify cognitive/perceptual ability constructs that tap abilities relatively independent of those measured by the ASVAB and that may be used to improve the Army's selection and classification decisions process.

Cognitive/perceptual ability constructs selected for inclusion in the Pilot Trial Battery and their designated priorities (in parentheses) are: (1) Spatial Visualization - Rotation and Scanning;

(2) Spatial Visualization - Field Independence; (3) Spatial Orientation; (4) Induction - Figural Reasoning; (5) Reaction Time - Processing Efficiency; (6) Memory - Number Operations; (7) Memory - Short Term Memory; (8) Perceptual Speed and Accuracy.

Determining the Method of Administration

In this section, we review the factors that influenced our decision to measure a particular construct via paper-and-pencil or via computer. The first factor concerns the construct definition and the dependent measures suggested by that definition. For example, definition of the construct, processing efficiency, indicates that the dependent measure involves the time required to respond to simple stimuli. Such information can only be obtained on a computer because a precise measure of reaction time is required. Hence, those constructs that involve a reaction time component, such as Processing Efficiency, Perceptual Speed and Accuracy, and Memory were slated for computer administration. McHenry and Toquam (1985) provide a detailed description of measures developed for computer administration.

The second factor involves the cost related to adapting items to the computer. For example, test items for such constructs as spatial visualization and figural reasoning involve detailed figures and objects. To adapt these items to the computer would require high resolution graphics. The cost for hardware capable of supporting such graphics at the time was prohibitive. Thus, we determined that measures of spatial visualization, spatial orientation, and induction would be assessed via paper-and-pencil. We focus on the development activities and pilot-test results of the new paper-and-pencil measures in the remainder of this paper.

Paper-and-Pencil Measures: Construct and Test Descriptions

In this section, we provide definitions of the constructs, describe criterion job performance areas or tasks that we expect measures of the constructs to predict and finally identify the tests designed to measure each construct. Detailed descriptions of the individual tests are available from the authors.

Spatial Visualization--Rotation

This involves the ability to mentally restructure or manipulate parts of a two- or three-dimensional figure. It serves as a potentially effective predictor of success in MOS that involve mechanical operations, construction and drawing or using maps. Two tests developed to measure this construct include Assembling Objects and Object Rotation.

Spatial Visualization--Scanning

This includes the ability to visually survey a complex field and to find a pathway through it. According to our expert judges, measures of this construct are potentially effective as predictors of success for Army MOS involving electrical or electronics operations, using maps in the field, and controlling air traffic. The two measures designed to assess this construct in the Path Test and the Maze Test.

Spatial Visualization--Field Independence

This includes the ability to find a simple form when it is hidden in a complex pattern. A measure of this construct is expected to predict success in MOS that involve detecting and identifying targets, using maps in the field, planning placement of tactical positions, air traffic control and troubleshooting operating systems. The Shapes Test was developed to measure this construct.

Spatial Orientation

This involves the ability to maintain one's bearing with respect to points on a compass and to maintain appreciation of one's location relative to landmarks in the environment. From job observations conducted in the field, we expect measures of this construct to predict success in combat MOS that involve maintaining directional orientation using features of landmarks in the environment. Three tests involving different orientation tasks were developed to assess this construct, Orientation 1, Orientation 2, and Orientation 3.

Induction - Figural Reasoning

This includes the ability to generate hypotheses about principles governing relationships among several objects. According to the panel of experts, measures of this construct are effective predictors of success in MOS involving troubleshooting, inspecting, and repairing electrical, mechanical, or electronic systems, analyzing data, controlling air traffic, and detecting and identifying targets. We developed two tests involving different tasks to assess abilities in this construct area. These were titled Reasoning 1 and Reasoning 2.

Test Development Issues

Two issues impacted on our approach for developing the new paper-and-pencil measures. These include the target population completing the new tests for selection and classification purposes and the power versus speed components of each new test. We discuss each in turn below. The population completing these tests is the same population that completes the ASVAB to qualify for entrance into to the Army. This is, very generally speaking, a population composed of predominantly recent high school graduates, not entering college, from all geographic sections of the United States. For our purposes the target population was, practically speaking, inaccessible during the test development phase. We were constrained to using enlisted soldiers to try out the newly developed tests. The development group, enlisted soldiers, of course, represents a restricted sample because they have passed enlistment standards and often have completed basic and advanced individual training.

Differences between the target population and the sample available to us, lead to two major implications that served as general guidelines for test development and pilot testing activities. First, the target population includes a broad range of abilities, therefore we attempted to develop test with a broad range of item difficulties. And second, the the test development group, first-term enlistees, would be of generally higher in ability than the target population. Therefore, the overall difficulty level of the test should be somewhat higher (i.e.,the test should be somewhat easier) than what it would have been if we had access to an unrestricted sample of the target population.

Another decision to be made about each test was its placement of the power vs. speed continuum. Most psychometricians would agree that a "pure" power test is a test administered such that all persons taking the test are allowed enough time to attempt all items on the test, and that a "pure" speeded test is a test administered such that no one or very few taking the test has enough time to attempt all items. In practice, there appears to be a power/speed continuum, most tests fall somewhere between the two extremes on this continuum.

During the preliminary test development stage, we categorized each test as a power test, speeded test, or combination of the two using our construct definitions. For example, using our definition of Induction, we designed the test items to represent a very wide range of difficulty levels and established a generous time limit such that most subjects would have time to complete all items. Thus, measures of induction were designed to fall on the power end of the continuum. Our plan for measures tapping Spatial Visualization -Rotation and Scanning differed from this in that all items were constructed to be moderately easy but more restrictive time limits were imposed. Thus, these measures were intended to fall toward the speeded end of the continuum.

For the remaining constructs, Spatial Visualization-Field Independence and Spatial Orientation, we designed the measures using the construct definitions to determine the range of item difficulties and to establish time limits. Following each pilot-test we examined completion rates and item difficulty levels to assess how closely performance on each new test matched the corresponding construct definition with regards to speed and power components.

Evaluating the Paper-and-Pencil Tests

Four pilot test or tryout sessions were conducted at Fort Carson, Fort Campbell, Fort Lewis, and Fort Knox. In the first pilot-test at Fort Carson, about 38 soldiers completed each paper-and-pencil test. The number at Fort Campbell was 57 and at Fort Lewis it was 118. At Fort Knox the numbers were 290 for time one and 97 to 126 for time two. Factors used to evaluate each test at one or more of these pilot-test sessions include the following: construct validity, test item characteristics, and test reliability. Below we present some general findings for all paper-and-pencil tests.

One goal of the the pilot-test sessions was to verify the construct validity of the new measures. Therefore, we identified published tests that measure constructs similar to our construct definitions. These published measures were included in the first three pilot-tests. It is important to note that, in general, most published tests or marker tests differed from the new tests in item difficulty levels and in the specific task required. Therefore, we did not expect a one-to-one correspondence between the new test and its published marker test.

Very few of the newly developed tests correlated above .65 with the designated marker; most correlations between new measures and marker tests fell between .45 and .60. These values were as expected given the differences in task requirements and in item difficulty levels between the new and marker tests. Basically this information suggested to us that although the tests did not duplicate their respective marker tests, they captured the essence of the target construct.

Another goal of the pilot-test sessions was to assess the psychometric characteristics of each new test. Following the pilot-test sessions, then, we computed item difficulty levels and item-total correlations for each test. These data were used to modify test items and to adjust time limits.

Results from the first pilot test indicated that all tests required some modification. For example, completion rates, item difficulty levels and raw total test scores suggested that some of the new measures may suffer from ceiling effects. Thus, for Assembling Objects, Object Rotation, Path Test, and Orientation 1, we constructed new items and adjusted time limits accordingly to obtain the desired difficulty level. For the Shapes Test and Maze Test, we modified test items to increase difficulty levels and to reduce the possibility of ceiling effects.

The reverse situation appeared on one of the orientation tests, Orientation 1. That is, item difficulty levels were low or the test was more difficult than desired. We modified this test by adding four "easy" items and by expanding the time limit.

For the remaining measures, Orientation 3, Reasoning 1, and Reasoning 2 very few changes were required. For example, item analysis data revealed that for some of the items, item-total correlations were higher for a distractor than for the correct response. These items were either modified or replaced.

Subsequent pilot tests indicated that the tests, in general, required only minor modifications.

Finally, we investigated the reliability or internal consistency and the stability of each new measure. To compute internal consistency estimates we used a split half procedure. This included administering each test as two separately timed halves and computing the correlation between part one and part two for each test. The Spearman-Brown correction procedure was then used to estimate the reliability for the test as a whole. We estimated the stability of each test by collecting test-retest data on a sample of about 100 soldiers at Fort Knox. A period of two weeks separated the two test sessions.

Internal consistency and test-retest estimates for each test appear in Table 1. Results from the Fort Lewis pilot-test indicate that the split half internal consistency estimates range from the high 70's to the low 90's for all tests with the exception of Reasoning 2. Test-retest estimates are lower than the internal consistency estimates but are at acceptable levels ranging from .57 to .84. The Reasoning 2 test once again yields the lowest value of all.

Note that in Table 1, we have also included internal consistency estimates for the Fort Knox sample computed using the Hoyt formula. and may represent overestimates for some of the more highly speeded tests. With the exception of Reasoning 2, these values range from the low 80's to high 90's.

Overlap Between the New Measures and the ASVAB

As we above, the major focus of this research involves identifying and developing measures of constructs not currently assessed in the ASVAB. One way to estimate the amount of overlap between each new measure and the measures contained in the ASVAB is to conduct uniqueness analyses. This procedure involves computing the squared multiple correlation between each new test and the ten ASVAB

TABLE 1

Reliability and Uniqueness Estimates for the Ten Paper-and-Pencil Tests Included in the Pilot Trial Battery

			Reliability			Uniqueness	
			Ft. Lewis	Ft. Knox			
		Time Allotted	r_{xx}		r_{xx}	ASVAB	ASVAB
Test	No. Items	(in minutes	Split Half N = 118	Alpha N = 290	Test-retest (N = 97 to 126)	R^2 Using Split Half	U^2 Using Split Half
Assembling Objects	40	16	.79	.92	.74	.40	.39
Object Rotation	90	7.5	.86	.97	.75	.19	.67
Mazes	24	5.5	.78	.89	.71	.25	.53
Path	44	8	.82	.92	.64	.29	.53
Shapes	54	16	.82	.92	.70	.19	.63
Reasoning 1	30	12	.78	.83	.64	.29	.49
Reasoning 2	32	10	.63	.65	.57	.26	.37
Orientation 1	150	10	.92	.98	.67	.36	.56
Orientation 2	24	10	.89	.88	.80	.30	.59
Orientation 3	20	12	.88	.90	.84	.54	.34

subtests. The resulting value is then subtracted from the reliable variance in that new measure (in this case we used the reliability estimate computed using the split half procedure). This value represents an index of the unique variance or variance that is uncorrelated with scores obtained on the ASVAB. Results from this analysis are reported in the final two columns in Table 1.

Across the ten new tests, the squared multiple correlations range from .54 to .19. It is clear that some of these tests are measuring abilities tapped by ASVAB subtests. On the other hand, the uniqueness estimates which range from .67 to .34, indicate that the new tests tap abilities independent from those assessed by the ASVAB subtests.

In sum, results from the uniqueness analysis are essentially what we would expect in assessing the amount of overlap between groups of tests that measure cognitive/perceptual abilities. The data are encouraging because they indicate that we are measuring ability constructs not currently assessed by the ASVAB.

REFERENCES

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1982). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10: 1981 Army applicant sample, (Technical Report 581). Alexandria, VA: U.S. Army Research Institute.

McHenry, J. J., & Toquam, J. L. (1985). Computerized assessment of perceptual and psychomotor abilities. Paper presented at the Military Testing Association Conference, 24 October, San Diego, CA.

Peterson, N. G. (1985). Mapping predictors to criterion space: Overview. Paper presented at the Military Testing Association Conference, 24 October, San Diego.

THREE VARIABLES THAT MAY INFLUENCE THE VALIDITY OF BIODATA

Clinton B. Walker
U.S. Army Research Institute

Presented on panel
"Utilization and Validation of Biodata"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

THREE VARIABLES THAT MAY INFLUENCE THE VALIDITY OF BIODATA[1]

Clinton B. Walker
U.S. Army Research Institute for the Behavioral and Social Sciences

This research examines the effect on predictive validity of traditional procedures for developing and implementing suitability screens in the military. For this paper, suitability screens in the form of background questionnaires, or biodata, will be considered. Typically, predictor tryouts have been run on new recruits whose subsequent performance has been tracked for the first six months of service (Atwater & Abrahams, 1983; Walker, 1985). Item selection and keying have then been based on the observed relation between predictor data and the criterion of successful service (versus discharge for bad causes). In the case of the U.S. Army's Military Applicant Profile (MAP), the instruments and keys have been implemented no less than two and a half years after the tryouts.

There is reason to suspect that three aspects of this traditional sequence - viz., testing recruits rather than applicants, tracking the cases for only six months, and implementing long after pilot testing - adversely affect operational validities. Since recruits and applicants are likely to differ in their desire to make themselves look good on self-report measures, applicants could be expected to try more than recruits to earn high scores. As a result, scoring keys that are developed on data from recruits may be less valid for scoring responses of applicants. In support of this hypothesis, Means and Heisey (1985) have found more self-serving responses in data from applicants than from recruits.

The hypothesis that using only a six-month tenure for tracking success/attrition lowers validities is based on the following two premises. First, more than half of attritions occur after the initial six months (Goodstadt & Yedlin, 1980; Hicks, 1981; Walker, 1985). Second, attrition during the first six months may not occur for the same reasons as later attrition. In the first six months recruits make their initial adjustment to military life while undergoing entry-level training; after that they are serving with operational units. Unfortunately, the archival codes for types of attrition are too cryptic (e.g., "Trainee Discharge Program," "Unsuitable Unknown," "In Lieu of Court Martial") to indicate whether earlier and later attrition are qualitatively different phenomena. But if they are, then using a longer than traditional criterion period for developing scoring keys might produce different keys.

A long lag time before implementing scoring keys is suspect because characteristics of the applicant pool change over time. Once the predictor data are collected for developing a biodata instrument, they may obsolesce as the criterion ripens. If the nature of the applicant pool changes much, then a scoring key may lose validity before it is ever used for screening, and continue to lose validity after implementation.

[1] Thanks go to Elizabeth P. Smith for advice on programming and to Winnie Young for creating the dataset on applicants in FY 81/82 from the Project A Longitudinal Research Database.

The present research uses data from the MAP to test the effects of each of these variables. Responses to a common set of items by contemporaneous applicants and recruits are compared to test the effect of examinees' status. Then, various statistics are examined over the course of years to test the effect of time lapse on the keys' validity. Finally, the effect of duration of the criterion period is tested by comparing the predictor responses of examinees who were discharged within and beyond the first six months of service. Each of those issues is treated in turn below in a separate section.

Applicants Versus Recruits

Method

To keep from confounding the effects of examinees' status with those of date of testing (i.e., temporal drift), it is necessary to compare contemporaneous applicants and recruits. Two such comparisons are available in the MAP data. First, MAP scores of 2,374 non-graduate applicants during FY 82 were compared with those of 1,286 non-graduate recruits who were tested in February-June, 1982. These recruits were the non-graduate subset of a sample of 9,603 cases on whom new instruments were being developed (Erwin, 1985). Out of the 240 items in that research, 38 were chosen for use here according to these two criteria: they had to be on the operational form of MAP, so the applicants would have taken them, and they must have shown validity for non-graduates in the developmental research. These 38 were the universe of items that met both criteria. The key for scoring had been developed on all 9,603 cases. Here the comparison was a *t*-test on the total score, 0 to 71 being the possible range.

Data for the second comparison overlap in part with the previous ones. In the developmental work of 1982, the item pool was administered to a sample of applicants at 39 Military Entrance Processing Stations (MEPS) nationwide and to recruits at all seven Army Reception Stations. Out of those groups, a respective 949 and 9,603 examinees of all levels of education, age, and gender were retained for analysis. Retention was based solely on the availability of individuals' criterion data in central personnel files. In the applicant sample, 267 cases retook the instrument later as members of the recruit sample. Presumably the presence of those cases reduces the between-group differences, thus biasing any test against finding differences.

The vehicle for this second comparison was two 101-item forms of MAP which were developed on the 9,603 recruits. These forms each had 78 unique items and 23 items in common, yielding possible scores of 1 to 188 on one and 0 to 194 on the other. Mean MAP scores and validities against the six-month tenure criterion were compared in the applicants and recruits.

Results

Descriptive statistics for the non-graduate applicants in FY 82 and the recruits in the 1982 development sample are included in Table 1. The observed difference in means of 10.9 points is significant ($t = 52.9$; $p < .001$) and the effect is strong (omega square = .43). Data for applicants and recruits in the 1982 developmental project are summarized in Table 2. Applicants' total scores were higher by 2.81 points on Form 1 and 2.1 points on Form 2. These differences gave *t*'s of 5.62 and 3.96 ($p < .01$ in each case). However, here the strength of effect was less than 1% for each form. For both forms, the observed validities were higher for recruits than for applicants. The difference between correlation coefficients for independent groups (Guilford & Fruchter, 1973) was computed on the validities for each form. The observed $\underline{z}$'s of 2.51 and 1.59 had one-tailed probabilities of .006 and .056, respectively.

Table 1
Descriptive statistics for four samples of non-graduates

Logical role	Date of predictor data	Status: Applic/ Recruit	n	r	Mean out of 0-71	SD	% finish 6 mo
Develop key	1-6/82	Recr	1,286	.18	33.3	6.2	79
X-validation & 0 yr drift	10/81-9/82	Appl	2,374	.02	44.2	5.5	86
1 yr drift	10/80-9/81	Appl	3,567	.07	44.2	5.2	86
2 yr drift	7/79-6/80	Appl	14,771	.01	28.3	5.6	86

The "instrument" for these data was 38 items from MAP 4B which were keyed on the total 1982 development sample of 9,603 cases and were also valid for its non-graduate subsample.

Table 2
Descriptive data for applicants and recruits in 1982 development sample

		Form 1			Form 2		
Status	n	Mean	SD	r	Mean	SD	r
Applicants	949	125.56	14.49	.24	123.72	15.54	.27
Recruits	9,603	122.75	16.64	.32	121.62	17.27	.32

Both samples include 267 cases who took the instrument a second time as members of the recruit sample.

Discussion

Both sets of comparisons support the hypotheses that applicants get significantly higher scores than recruits, even though both samples were selected on the basis of operational MAP. Although the comparison of validities favors the hypothesis, that evidence is weakened by the fact that the recruit sample was also the sample on which the scoring key was developed. Nevertheless, the generalizability of data from recruits to applicants is not supported here.

Drift in Validity

Method

For examining possible loss of validity over time, a non-operational key was used that had been developed on the 1982 recruit data. The criterion was successful completion of the first six months of service (vs. discharge for failures to adapt). The "instrument" consisted of the 38 items mentioned earlier. Meeting the criteria of being on the operational form of MAP and being validated on non-graduates, the items could be used to compare results for non-graduates in different year groups who took MAP before entering the service. Three samples of such applicants were available: 2,374 in FY 82, 3,567 in FY 81, and 14,771 in 7/79-6/80. Because the 1982 key was not cross-validated by the developer, the 1982 applicants became a cross-validation sample. Thus, their data were used to see how much validity there was to drift in the first place. Validities in the form of Pearson *r*'s, mean total scores for the 38 items, and success rates (i.e., percent of sample completing the first six months of service) were compared over the four samples.

Results

Table 1 gives descriptive statistics for the recruits in 1982 and for three samples of applicants. In contrast with the original value of .18, validities for applicants in 1982, 1981, and 1979/80 were .02, .07, and .01, in order. The key did not effectively discriminate between examinees who went on to complete the first six months of service and those who did not: mean differences in scores for those two criterion groups reached a maximum of .18 *SD* in the three samples. Means out of a possible 71 points ranged from 28.3 to 44.2 points in the four groups, while success rates varied from .79 to .86. Using the 1982 applicants as a basis for confidence intervals on the means, we find significant differences ($p < .001$) in both the 1979/80 applicants and in the 1982 development sample. The normal approximation to the binomial found the development sample to have a significantly lower attrition rate than the 1982 applicants ($z = 12.28$; $p < .001$), all of whom had entered the Army.

Discussion

The low validity that was observed in the 1982 applicants amounts to a failure of the (non-operational) 1982 key to cross-validate. Thus, there was little if any original validity that could drift. Absent drift in validity, however, there were significant jumps in both predictor and criterion scores across samples. If changes occur in validity over time, they could be due to gradual trends in the population of applicants, to short range instability in the population, or to both. It is possible that similar variability could be found in subsamples of the 1982 recruits. In order for the 1982 developmental data to have any hope of producing a durable key, they would have to undergo a legitimate cross-validation. Elizabeth P. Smith and I are now working on this problem in-house at the Army Research Institute.

Six Months Versus Longer Tenure

Method

An operational form of MAP, Form 4B, gave the data for this analysis. Its 60 multiple choice items were validated in 1977 on 2,280 male recruits who had not completed high school (Frank & Erwin, 1978). In content, the questions cover experiences in school, extracurricular activities, work history, and expectations of life in the service. The present examinees were 2,564 17-year old non-graduate males. They all took MAP as a pre-induction screen in Fiscal Years 81/82, entered the Army, and then received adverse discharges in their first tour. For the first analysis, examinees were split into two groups, those discharged within the first six months of service ($\underline{n}$=860) and those with longer tenures ($\bar{x}$=363 days; $\underline{n}$=1,704). For each of the 60 items, a chi-square test of association was run on frequencies of response for each alternative by group. Cramer's $\underline{V}$ for the items was examined as well for estimates of strength of effects. To judge the potential of response choices for keying, differences between groups in rates of endorsing individual choices were examined in items giving a significant groups-by-response choice chi square.

A second similar analysis was done to see whether the sensitivity of biodata items to individual differences in adaptability is masked by lumping successful cases with those who receive bad discharges after six months. For this analysis, chi-square tests were run twice on the total sample of 5,941 non-graduate applicants in FY 81/82. This sample included those who served successfully. The sample was split differently for these runs: once as all dischargees vs. all successful cases, and once as all discharges within six months of entry vs. all other cases. Simple numbers of significant ($\underline{p}$ < .05) chi squares and median $\underline{p}$ values from the two splits were compared.

Results

In the analysis of dischargees, 10 of the 60 group-by-response choice chi square tests gave probabilities < .05. Of those, three had $\underline{p}$'s < .01. Cramer's $\underline{V}$ for the ten items ranged from .056 to .085, while $\underline{V}$'s for seven items with .05 < $\underline{p}$ < .15 were also above .05. The median level of significance for all 60 items was .30. In each of the ten items with the lowest $\underline{p}$ values, the single response choice which had the greatest difference between groups in rate of endorsement was tallied. The median of those ten maximal differences was 4.2% (range: 3.19 - 6.78%).

In the second analysis, 13 of the chi-squares on items gave $\underline{p}$ < .01 when the positive criterion group included bad discharges after six months. In contrast, when the criteron groups are pure (i.e., all bad discharges vs. only the successful cases), the significant items rise to 25. Median $\underline{p}$ values under the two conditions are .31 and .15, in order.

Discussion

The differences in response distributions are small for examinees who were discharged before and after six months. Given that the significance of chi-square is inflated by large sample sizes, and that the probability of Type I errors is great in such a large set of significance tests, a finding of ten items out of sixty with $\underline{p}$ < .05 is not large. Also, given the small values of $\underline{V}$ for those ten items and the small between-group differences in response frequencies, the data do not support keying the instrument separately for the periods of initial and field service. As for causes of attrition, the very similar distributions of predictor responses for the two groups in this dataset do not imply that the reasons for early and late attrition differ.

Although the usefulness of keying long and short tenures differently is not supported, the value of using a longer criterion tenure for key development is. In the analyses here, almost twice as many items were sensitive to real differences in success when the positive criterion group was purged of later attritions. The practice of developing keys on six month success seems here to undermine the validity of the predictor.

Conclusions

We now have evidence that traditional practices in developing biodata may have major flaws. A system for countering these problems is easy to conceive. Starting with a validated instrument, we could continually gather predictor scores of applicants and criterion scores of accessions. Today's selection measures would also be used as the predictor data for a later generation of scoring key, which would be based also on the performance measures. Updating of keys would then be ongoing rather than rare and ad hoc, as it is now. With ongoing updating, keys would be available after a minimal time lag and with appropriate generalizability (i.e., from applicants to applicants). Of course, increasing the criterion tenure would increase the time until new keys were available, but the best tradeoff between lag and quality could be determined empirically. Although problems in operating a biodata screen have been documented here, practical solutions are available.

References

Atwater, D. C. & Abrahams, N. M. (1983). Adaptability screening: Development and initial validation of the Recruiting Background Questionnaire (RBQ) (NPRDC TR-84-11). San Diego: Navy Personnel Research and Development Center.

Erwin, F. W. (1985). Development of new Military Applicant Profile (MAP) biographical questionnaires for use in predicting early Army attrition. Unpublished manuscript.

Frank, B. A. & Erwin, F. W. (1978). The prediction of early Army Attrition through the use of autobiographical information questionnaires (Technical Report No. TR-78-A11). Alexandria, VA: Army Research Institute.

Guilford, J. P. & Fruchter, B. (1973). Fundamental statistics in psychology and educaton. New York: McGraw-Hill.

Goodstat, B. E. & Yedlin, N. C. (1980). First tour attrition: implications for policy and research (Research Report 1246). Ft. Benjamin Harrison, IN: Army Research Institute.

Hicks, J. M. (1981, March). Trends in first-tour armed services enlisted attrition rates. Paper presented at the annual meetings of the Southeastern Psychological Association. Atlanta, GA.

Means, B. & Heisey, J. (1985, in press). Educational and biographical data as predictors of early attrition. Alexandria, VA: Human Resources Research Organization.

Walker, C. B. (1985). The Army's Military Applicant Profile: Its background and progress. In B. Means (Editor), Recent developments in military suitability research. (In preparation)

LEADERS' BEHAVIOR AND THE PERFORMANCE OF FIRST TERM SOLDIERS

Leonard A. White
Ilene F. Gast
Michael G. Rumsey

U.S. Army Research Institute

Presented on panel,
"Dimensions of Leadership"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

LEADERS' BEHAVIOR AND THE PERFORMANCE OF FIRST TERM SOLDIERS

Leonard A. White, Ilene F. Gast and Michael G. Rumsey[1]
U.S. Army Research Institute for the Behavioral and Social Sciences

A large Army project is underway to validate new and current predictors of first term soldier performance. A major objective of this effort is to increase Army organizational effectiveness by improving the soldier job match. This will be accomplished by developing a set of selection and classification measures (predictors) and performance criteria and then empirically demonstrating relationships between the predictors and performance measures.

However, soldiers' performance on the job is not only related to the personal characteristics which they have, but to experiences and developmental opportunities throughout their life-cycle in the Army. Longitudinal research indicates that the quality of leader-subordinate work relationships are predictive of job success (Wakabayashi & Graen, 1984). Aspects of leader behavior such as providing rewards and recognition, disciplinary practices, and inspirational leadership have been related to subordinates' effort and performance (e.g., Yukl, 1981).

Past research on leadership and performance has generally omitted the influence of ability or the potential interactive effect between individual aptitudes and leadership on job proficiency and performance. Some investigations (e.g., Barnes, Potter, & Fiedler, 1983) have suggested that the prediction of job performance from general ability is moderated by leadership. Other researchers (Schmidt & Hunter, 1977) have argued that the relationship between general ability and performance is stable across time and situations for similar jobs.

To summarize, the model examined in this research assumes that job performance is influenced by a new incumbent's capabilities measured prior to enlistment and characteristics of the work environment. Within this framework the present research uses data from Project A to: (a) examine relationships among leader actions and subordinate performance, and (b) to explore possible moderating effects of leadership on the correlation between general cognitive ability and job performance.

METHOD

Research participants were 696 first term soldiers in five military occupational specialties (MOS); 156 infantrymen (MOS 11B), 139 armor crewmen (MOS 19E), 125 radio teletype operators (MOS 31C), 141 light wheel vehicle mechanics (MOS 63B), and 135 medical care specialists (MOS 91A).

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

Of these soldiers, 88.5% were male and 11.5% were female; 28% were black, 3% were hispanic, 64% were white, and 5% other. Soldiers' report of work experience in their unit ranged from 2 months to 49 months (median=one year).

Instruments

The first step in this research was to develop measures of leader behavior and soldier performance on the job.

Supervisor behavior rating scales. Critical incidents workshops were conducted with 80 NCO in the five target MOS. These NCO generated a total of 474 examples of leader behaviors thought to influence soldier performance. Classification of the incidents by two of the authors and 31 NCO familiar with Army leadership requirements led to the identification of 9 categories of leader behavior (White, Gast, Sperling, & Rumsey, 1984). At least 5 and no more than 8 items were written to represent important leader behaviors in each category (e.g., Your supervisors are hard to find when you need them). These procedures resulted in a 60-item question naire. Responses to each item were made on a 5-point scale from very seldom or never (1) to very often or always (5).

Job performance rating scales. To develop these scales, critical incident workshops were conducted in which NCO provided examples of effective (as well as ineffective) soldier performance. The number of NCO and examples provided were as follows: MOS 11B, 51 NCO's, and 906 incidents; MOS 19E, 43 NCO's and 798 examples; MOS 31C, 45 NCO's and 830 incidents; MOS 63B, 49 NCO's and 882 incidents and; MOS 91A, 42 NCO's and 783 incidents. A variant of the behaviorally anchored rating procedure (Smith & Kendall, 1963) was used to develop behavior-based rating scales for each job. The resulting rating form for each job consisted of seven to ten 7-point behavior summary scales.

Army-wide performance rating scales. To prepare these scales, 77 NCO's and junior officers working in a wide variety of Army jobs generated 1,215 behavioral examples. The examples represent those aspects of soldier effectiveness that contribute, broadly speaking, to organizational effectiveness, such as following orders and regulations. The target criterion space for these scales went beyond job performance to include aspects of socialization and commitment to the organization. Eleven 7-point behavior-based rating scales were developed for each job.

Hands-on, task proficiency tests. For each of the jobs, 5-8 critical tasks were identified to represent the MOS-specific task domain. Multistep task proficiency tests were prepared for each task. Each step of a task was scored pass or fail. A score for each task was computed by calculating the proportion of steps passed and the task scores were averaged to yield an overall hands-on test score.

Job knowledge tests. Through job analysis, important knowledge areas were identified for each of the five jobs. With the help of subject matter experts, items were written to tap these knowledges. For each soldier, the percentage of correct items was the overall job knowledge test score.

General cognitive ability. The Armed Services Vocational Aptitude Battery (ASVAB) was administered to all participating soldiers prior to entering military service. The ASVAB, which consists of ten subtests, is used for selection and occupational classification. A composite measure of four ASVAB subtests, known as the Armed Forces Qualification Test (AFQT), was used as the measure of general cognitive ability.

Procedure

Raters were trained to use the behavior-based rating scales. After training, supervisors in groups of 3-15 evaluated their subordinates on the Army-wide and job performance rating scales. The mean number of supervisor raters/ratee ranged from 1.66-1.83 for the five MOS. Ratings were averaged across supervisor raters to form an overall job performance rating and an Army-wide effectiveness rating for each ratee.

The first term soldier (ratees) completed the supervisor behavior rating scales, and were also administered tests of job knowledge and hands-on, task proficiencies.

RESULTS

Principal components factor analysis was used to examine the dimensionality of the supervisor behavior rating scales. Varimax and promax solutions were computed and the interpreta tion restricted to factors appearing in both solutions. Comparison of the rotated structures yielded eight factors with eigenvalues greater than one. Items loading above .4 on one and only one factor were interpreted as measuring the factor. Items with weak loadings on all factors or similar loadings on two or more factors were not used to measure any factor. Factor score estimates were computed by unit weighting and summing individual's responses to the set of items representing each factor. Table 1 presents the intercorrelations among the estimated factor scores.

Table 1

Intercorrelations Among Leadership Scales and Summary Statistics.

Scale	1	2	3	4	5	6	7	8	9	No. of Items	Scale Mean	Std. Dev.
1. Support/Inspiration	.89	.64	.48	.64	.53	.72	.58	.68	.90	9	27.2	7.5
2. Informing		.78	.54	.49	.53	.54	.48	.50	.79	6	19.0	4.8
3. Fairness			.74	.47	.46	.40	.31	.36	.67	5	16.5	4.3
4. Participation				.70	.44	.60	.46	.56	.76	4	13.4	3.4
5. Performance Contingencies					.55	.47	.40	.39	.67	3	9.9	2.5
6. Role Clarification						.78	.55	.63	.80	4	12.9	3.1
7. Results Orientation							.56	.59	.66	3	9.4	2.2
8. Training and Development								.72	.77	5	14.7	3.9
9. Total									.94	39	123.1	24.7

Note. Internal consistency reliabilities are presented on the diagonal.

[a] N = 696

Correlations of hands-on and job knowledge test scores, job performance ratings, and the Army-wide effectiveness rating with the leader behavior scales are presented in Table 2. Results are shown separately for each of the five jobs. A mean correlation ($\bar{r}$) across the five jobs was computed by weighting each correlation by its associated sample size (Hunter, Schmidt, & Jackson, 1982). The highest correlations were

Table 2

Correlations between Leadership Scales and Criterion Measures by Army Job

	Leadership Scale								
Job	1	2	3	4	5	6	7	8	Total Scale
				Hands-on Task Proficiency					
11B	.05	.03	.19*	.24*	.17*	.25*	.10	.11	.18*
19E	.14	.11	-.01	.14	.24*	.11	.18	.23*	.22*
31C	.02	.00	.15*	.18*	.02	.08	.03	.15	.09
63B	.12	-.05	.10	.06	.00	.05	.08	.12	.09
91A	-.05	-.14	-.12	.02	-.04	-.08	-.09	-.08	-.10
$\bar{r}$	.05	-.02	.08	.13*	.07	.09*	.05	.10*	.09*
				Job Knowledge					
11B	-.01	-.17*	.02	.13	.03	.15*	.09	.12	.03
19E	-.03	-.03	.05	.03	.06	-.13	-.02	-.03	-.02
31C	.17	.11	.12	.30*	.26*	.23*	.12	.17	.22*
63B	-.05	-.04	.05	.05	.20*	-.05	.01	-.06	-.01
91A	-.12	-.10	-.01	-.01	-.11	-.18*	-.22*	-.23*	-.13
$\bar{r}$	-.01	-.06	.04	.09*	.08	.00	.00	-.01	.01
				Job Performance Rating					
11B	.12	.01	.05	.23*	.08	.21*	.10	.03	.13
19E	.11	.04	.09	.16	.08	.05	.11	.21*	.13
31C	.21*	.01	-.04	.12	.20*	.17	.02	.07	.14
63B	.17	.08	.23*	.08	.20*	.07	.07	.08	.18*
91A	.08	.07	.05	.11	.00	.08	-.12	.01	.03
$\bar{r}$	.13*	.04	.08	.14*	.11*	.11*	.04	.08	.12*
				Army-Wide Performance Rating					
11B	.17*	.06	.04	.23*	.12	.20*	.11	.07	.17*
19E	.12	.02	.07	.14	.22*	.10	.14	.15	.15*
31C	.41*	.19*	.12	.34*	.32*	.32*	.18*	.24*	.37*
63B	.20*	.13	.30*	.11	.19*	.06	.04	.08	.21*
91A	.17*	.05	.14	.20*	.07	.09	-.09	.07	.12
$\bar{r}$	.21*	.09*	.13*	.20*	.18*	.15*	.07	.12*	.20*

Note. Leadership scales: 1 (Support); 2 (Informing); 3 (Fairness); 4 (Participation); 5 (Performance Contingencies); 6 (Role Clarification); 7 (Results Orientation); 8 (Training & Development); 9 (Total).

*$p < .05$

obtained between perceptions of leader behavior and the Army-wide effectiveness ratings. Within the set of Army-wide performance dimensions, strongest relationships were obtained between supportive and participative leadership and ratings of subordinate adherence to regulations and willingness to provide extra effort when needed. Statistically significant but low correlations between leader behaviors and job proficiency were evident in the two combat MOS.

Hierarchial regression analysis was used to estimate the relationships of cognitive ability (i.e. AFQT score), leadership climate, and their interaction to job proficiency and performance. The AFQT score was entered first in the regression to assess the contribution of mental ability at the time of enlistment to later job performance. Then, leadership and the ability X leadership interaction were entered to assess post-enlistment leader influences on performance and the utilization of ability on the job. In the regression model, leadership was represented by the sum of scores on the 8 leadership scales. The criterion variables were job knowledge, hands on task proficiency, and supervisor ratings of job performance and Army-wide effectiveness.

Of interest here, results of the regression analyses revealed no statistically significant increase in $\underline{R}^2$ due to inclusion of the ability X leadership interaction in the model. In each of the five jobs, the highest multiple correlations were obtained for prediction of job knowledge, with $\underline{R}$=.30, to .60, all $\underline{p}$<.05. This effect was primarily attributable to the influence of general ability on job knowledge. Leadership and cognitive ability had significant independent effects on task proficiency in the infantryman and armor crewman jobs with, respectively, $\underline{R}$=.28, p<.05, and $\underline{R}$=.37, p<.05. However, in MOS 91A and MOS 63B $\underline{R}^2$s for the prediction of task proficiency from the independent variables failed to reach significance. With respect to supervisory ratings of job performance, ability and leadership and their interaction accounted for less than 5% of the variance in this criterion. Leadership showed several significant correlations with Army-wide effectiveness ratings at the zero-order level, however the $\underline{R}^2$ for this criterion achieved significance only in the radio-teletype operator job, with $\underline{R}$=.37, $\underline{p}$<.05. Correlations between cognitive ability and the Army-wide effectiveness rating ranged from $\underline{r}$= -.28 to .03.

DISCUSSION

The present research explored relationships between leadership, cognitive ability, and the performance of first term enlisted soldiers. Results for the five Army jobs examined here support the conclusion that general ability and leadership behavior have independent effects on performance. However, each appears to contribute to effective soldiering in different ways. Leadership, as perceived by the subordinate, had the strongest effect on the motivation-related, dependability facets of performance measured by the behaviorally based rating scales. General cognitive ability contributed to performance by enabling enlistees to learn the facts and procedures required to perform their jobs.

No evidence was obtained indicating that relationships between general ability and job proficiency and performance are moderated by leadership influences. This finding supports conclusions by Schmidt and

Hunter (1977) that the validities of cognitive tests are similar across situations for the same job. Correlations between general cognitive ability and each criterion measure did vary somewhat across jobs, but almost all of the variation was attributable to sampling error.

The relationships between leadership and performance reported here should not be interpreted as indicating that leadership behavior "causes" performance. Leadership effects on performance may be understood in terms of exchange theory (Graen, 1976) which views the interaction between leader and subordinate as a reciprocal influence process that develops over time. Subordinates who are perceived as willing to work hard and support the mission will be evaluated more favorably by their superiors. In return for their support, these soldiers are likely to receive more individualized attention, information, and other resources from their supervisors; which, in turn, serves to reinforce and sustain subordinate effort.

The results reported here are largely exploratory. Future data collection and analysis will provide an opportunity to confirm the leadership factors and to examine potential moderating effects of leadership behavior on a broad range of soldier aptitudes and characteristics.

REFERENCES

Barnes, V., Potter, E. H., & Fiedler, F. E. (1983). Effect of interpersonal stress on prediction of academic performance. Journal of Applied Psychology, 68, 686-697.

Graen, G. (1976). Role-making processes within complex organizations. In M. Dunnette (ed.) Handbook of Industrial Organizational Psychology. Chicago: Rand McNally.

Hunter, J.E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating results across studies. Beverly Hills: Sage Publications.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Wakabayashi, M., & Graen, G. B. (1984). The Japanese career progress study: A 7-year follow-up. Journal of Applied Psychology, 69, 603-614.

White, L. A., Gast, I. F., Sperling, H. M., & Rumsey, M. G. (1984, Oct). Influence of soldiers' experiences with supervisors on performance during the first tour. Paper presented at meeting of the Military Testing Association, Munich, Germany.

Yukl, G. A. (1981). Leadership in organizations. Englewood Cliffs, NJ: Prentice-Hall.

VOCATIONAL INTERESTS AS PREDICTORS OF ARMY PERFORMANCE

Hilda Wing
U.S. Army Research Institute

Bruce N. Barge and Leaetta M. Hough
Personnel Decisions Research Institute

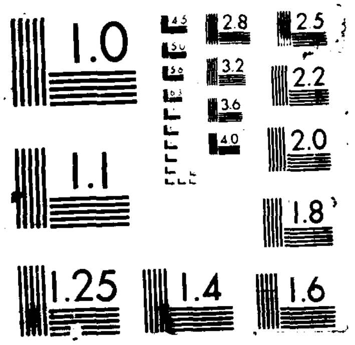
Presented on symposium,
"Elements of a Military Occupational Exploration System"

At the Annual Conference of the
Military Testing Association
San Diego, California

October 1985

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

AD-A198 274 IMPROVING THE SELECTION CLASSIFICATION AND UTILIZATION 2/5
OF ARMY ENLISTED P (U) HUMAN RESOURCES RESEARCH
ORGANIZATION ALEXANDRIA VA J P CAMPBELL MAY 88
UNCLASSIFIED HUMRRO-IR-PRD-87-12-SUPPL F/G 5/9 NL


1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

Vocational Interests as Predictors of Army Performance[1]

Hilda Wing
U.S. Army Research Institute

Bruce N. Barge and Leaetta M. Hough
Personnel Decisions Research Institute

In: Elements of a Military Occupational Exploration System
Military Testing Association, October 1985
San Diego, California

Measures of vocational and occupational interest have been used in selection for Army enlisted occupations for many years. In this paper we will describe how such measures have been used in the recent past, review current Army research which will link such measures to performance in Army jobs, and identify critical issues that must be resolved in order for interest measures to be effective in a selection and classification program.

The Army's Use of Interest Measures in Selection/Classification

Use of vocational interest measures for classification into Army training was part of the Army's selection and classification for enlisted personnel from 1958 until 1980. The Army Classification Batteries, followed by the Armed Services Vocational Aptitude Battery (ASVAB) Forms 6 and 7, included forms of the Army Classification Inventory (ACI), which contained sentences describing activities with which an applicant could agree or disagree. Four scale-scores were obtained from each applicant: Combat, Administrative, Mechanical, Electronics. These scale-scores were incorporated with ASVAB cognitive ability subtest scores to produce Aptitude Area (AA) composites. For example, the Combat AA included both ability and interest measures. Empirical data supporting this use had been provided by the developers of ACB-73 (Maier & Fuchs, 1972). Interest measures were dropped from the Army enlisted classification system with the introduction of new ASVAB forms in October of 1980.

The Army's current Project A is, among other things, the largest selection and classification research effort to date. The initial function of Project A is to validate the ASVAB against Army performance. An additional aspect of Project A's mission is to develop new predictors which will cover attributes that the ASVAB does not. ASVAB is more than adequate for selection into Army training (McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984). What we are more concerned about is classification and, in addition, performance on the job, successful completion of the first tour, and reenlistment eligibility. To that end there has been

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Deparment of the Army.

developed an evolutionary model of predictor space. This model conceives of predictor space as having three components. First, the cognitive-perceptual component includes measures of verbal, quantitative, and spatial abilities. Next, the perceptual-psychomotor component includes perceptual speed and accuracy, short-term memory, multi-limb coordination, and movement judgment. We have developed a mini-battery for this component which is administered on an IBM-compatible microcomputer with a custom-designed response pedestal. Finally, the non-cognitive component covers both biographical/temperament (personality) and vocational interest measures.

The evolution of this model of predictor space has been firmly anchored to data, as follows. Project A has so far completed research on a first cohort of Army enlisted personnel, those entering in FY 1981 and 1982. The second cohort includes those soldiers who entered the enlisted service during FY 1983-1984. It includes both longitudinal and concurrent components; the longitudinal is included in the concurrent. For the longitudinal effort, we developed our first test battery, the Preliminary Battery, from readily available, off-the-shelf paper and pencil measures of cognitive and non-cognitive attributes. We administered the Preliminary Battery prior to training, to soldiers in four selected MOS, from October 1983 through June 1984. This year we obtained measures of training success and early attrition for this sample.

This summer, we are testing the concurrent component of this 1983-84 cohort with a second, new battery, in conjunction with a full complement of performance measures. We have added another 15 MOS, and the perceptual-psychomotor component of predictor space is being evaluated with micro-computers. Data collection should be complete by late November, 1985. While we have no analyses completed for what we are calling the Trial Battery, we do have some information about its immediate fore-runner, the Pilot Trial Battery.

A complete longitudinal effort is planned for the FY 1986-1987 cohort, to begin sometime next year. There will be the Experimental Battery, which will be much like the Trial Battery, to be administered to soldiers entering training in each of our selected MOS. Subsequently, we will be administering the appropriate performance measures to these soldiers. At the same time, we also plan to evaluate the performance of second-tour members of our 1983-1984 cohort.

Results for the Preliminary Battery

The Preliminary Battery included the Air Force Vocational Interest Career Examination (VOICE), which assesses 18 basic interests (Alley & Matthews, 1982). Because of the research on the Holland hexagonal model of vocational interests, we investigated its appropriateness. We factor-analyzed both the items and scales of the VOICE. We were able to recover the 18 basic interest scales quite nicely from the item factor analyses (Hough, Dunnette, Wing, Houston, & Peterson, 1984). We were able to find the Realistic group of occupational interests, but all the others clumped mostly into one group. In hindsight this made perfect sense. The majority of occupations in the enlisted military service are Realistic in nature, as they are jobs in the skilled trades. There are a handful of Investigative occupations, some Conventional, and some Social occupations. For virtually no occupation in the enlisted ranks does the Artistic or Enterprising label fit.

What evidence was there of criterion-related validity for these interest measures? Available criteria were existing training grades and early attrition (status as of December 1984, or an average of one year of service). For training, the cognitive tests of the Preliminary Battery appeared to have some predictive power, although the coefficients were not large and not much larger than those obtained for the ASVAB. The attrition analyses are currently incomplete. This criterion will be especially hard to predict because the early attrition was fairly low, about eight percent. While some of the VOICE scales were significantly related to attrition in each of the four MOS, the correlations were quite low. The coefficients for some of the biodata/temperament scales, which evaluated aspects of socialization, were higher than those for interests. The domain of causes for discharge in the Army extends from "disciplinary" through "for good of service" to "unsuitable unknown." It is likely that early attrition in the Army, particularly that through the Trainee Discharge program, may be more disciplinary than anything else. Thus, the predictiveness of the socialization scales is understandable.

The VOICE scales were not related to any great extent with the other measures evaluated, including the ASVAB. It is likely that as various criteria mature (later attrition, re-enlistment) or are administered as part of the Project A data collection (commitment, effectiveness), these early measures of vocational and occupational interests will have a better chance to demonstrate what they can do.

Results from the Pilot Trial Battery

The Pilot Trial Battery was field tested during the fall of 1984. Soldiers supplied data to evaluate the properties of the battery, including test-retest stability. We called our interest measure here the "Army VOICE," or AVOICE. We obtained this by starting with the VOICE, cutting back items on most of the 18 scales while adding scales for Army interests which are not duplicated in the Air Force, such as Infantry, Armor/Cannon, Science/Chemical Operations.

Psychometrically, the new instrument worked well, except that the factor analyses yielded the same pair of factors as before. For the Pilot Trial Battery, these factors appeared to be described better as "Combat" and "Combat Support," rather than "Realistic" and "Non-Realistic." This is a matter of taste rather than substance, as there is confounding of terms. The Combat occupations are Realistic while the Combat Support occupations cover the other five corners of Holland's hexagon. But, this is, we judge, the occupational reality of the Army enlisted world. The reliability and stability of the interest scales were excellent, in the .80's and .90's. There were no performance criteria available for this sample, but we did inspect the overlap of the interest measures with the remaining components of the Pilot Trial Battery and the ASVAB. The intercorrelations between AVOICE scales and other scales were generally low.

Preparing the Trial Battery from the Pilot Trial Battery consisted mainly of cutting back, so that a 6-7 hour battery was reduced to one requiring less than four hours. The AVOICE in the Trial Battery being administered now includes 176 items and takes about 15-20 minutes to administer. It will provide scores for interests in 22 Army occupations.

Issues in the Operational Use of Interest Measures in Selection and Classification

We see at least five major issues to be confronted in determining when and how to use measures of vocational interests in selecting and classifying for military enlistment. The first four are clearly technical while the last is more of a policy issue which can be informed by our technology.

First, the complete hexagonal model of Holland's vocational interest theory appears to be inappropriate for predicting performance in Army occupations. We tend to forget the context-sensitivity of models. The domain of Army jobs maps onto only a portion on the theorized hexagonal interest space, mainly that corner called Realistic. All of the other Army jobs, which could be characterized as involving interests from the Investigative, Social, and Conventional corners, appear to clump together. At this time we do not know whether this simple differentiation will provide all the predictability possible, given the available criteria, or whether further distinction into occupational scales will be warranted. But, it is clear that approaches using a complete Holland model will have limited applicability for the spectrum of Army enlisted occupations.

Second, the selection of appropriate criteria for vocational interests to predict is a major concern. Should criteria be those we consider as maximal effort, such as job knowledge tests and hands-on measures? Or should they be typical effort types of measures, such as motivation? We really need to know more about these criteria. One of the goals of Project A is to improve our conceptual understanding of the criterion space. This is clearly a worthy and necessary goal.

Third, how should predictors and criteria be used? The primary function of any interest measure is to direct the individual towards some occupations and away from others. That is, the object is classification. Regardless of the specific criteria, there are questions about the form of the predictors to use. Should we use scores from occupational scales, or should we use factor scores? Should we use single scores, or do we need to investigate configurations, or profiles? How should we combine interest measures with measures from other domains, such as the cognitive? It could be that positive interest in a specific area can compensate, to some extent, for lower ability for that area (Matthews, 1982). What are the characteristics of the sample sizes, the psychometric properties of the measures, that must be present for us to be able to make any kind of definitive statement concerning such claims?

Fourth, what exactly are we trying to predict: Success or avoidance of failure? This is the more complex issue concerning the fact that the

Army, and perhaps most employers in general, cannot always use people in what those people are best at. For example, one of the MOS in Project A is the Combat Medic. We have administered a complete battery of performance measures to several hundred Combat Medics so far in addition to the Trial Battery. However, at this point in time the United States is currently not in any general armed conflict, and there is little opportunity for these soldiers to practice their training in any realistic environment. Some of them may be working in maternity wards while others spend most of their time in the motor pool. We find it difficult to understand exactly how an interest in medical activities, absent other information, will be predictive of important criteria for these soldiers. Other examples are possible. How should interest measures be used in such cases?

The fifth and final issue concerns where in the enlistment process is it most appropriate to use interest measures? In the All-Volunteer Army, they may be more appropriately used by the recruiter. Should they be used in a mandatory or advisory way? Perhaps this is a technical question as much as are the other four: Are interest measures more predictive, of whatever criteria we can come up with, in whatever psychometric fashion determined effective, when these measures are used in an advisory fashion rather than a mandatory one?

This report has provided a brief description of how the Army is investigating the use of vocational interests in predicting performance in Army jobs. Project A will be providing vast amounts of data which will better inform our use of these measures. However, this use may be complex. The empirical data will, we trust, point us towards better use.

References

Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination: A description of the instrument and possible implications. Journal of Psychology, 112, 169-193.

Hough, L. M., Dunnette, M. D., Wing, H., Houston, J., & Peterson, N. G. (1984). Covariance analyses of cognitive and noncognitive measures in Army recruits. Paper presented at the convention of the American Psychological Association, Toronto, Ontario, Canada.

Maier, M. H., & Fuchs, E. F. (1972). Development and evaluation of a new ACB and aptitude area system (Technical Research Note 239). Alexandria, VA: U.S. Army Research Institute.

Matthews, M. D. (1982). Vocational interests, job satisfaction, and turnover among Air Force enlistees. Paper presented at the Fourth Annual Learning Technology Congress and Exposition, Society for Applied Learning Technology, Orlando, FL.

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative Armed Services Vocational Aptitude Battery (ASVAB) area composites (Technical Report 651). Alexandria, VA: U. S. Army Research Institute.

PERFORMANCE CRITERION MEASUREMENT:
WHAT ARE THE DIFFERENT METHODS MEASURING?

Walter C. Borman
Personnel Decisions Research Institute

Presented at the
Air Force Conference on Job Performance Measurement
San Antonio, Texas

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Performance Criterion Measurement:
What are the Different Methods Measuring?

Walter C. Borman

Personnel Decisions Research Institute
43 Main Street Southeast, Suite 405
Minneapolis, Minnesota 55414

INTRODUCTION

The Army Research Institute for the Behavioral and Social Sciences (ARI) initiated Project A, a nine-year research program intended to link selection and classification standards to job performance. The primary goal of Project A is to achieve increased Army effectiveness through improving the soldier-job match. This goal will be accomplished by developing a comprehensive set of selection and classification measures (predictors) and performance criteria, and empirically investigating correlations between these predictor and performance measures.

This paper explores relationships between different kinds of criterion measures in a large sample (N = 5021) of first-term soldiers in nine Army jobs. Performance rating scales, hands-on task proficiency measures, and job knowledge tests were all developed in Project A and administered during this large-scale concurrent validation (CV) data collection. Relationships between scores on these different criterion measures and between criterion

The research reported was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract MDA 903-82-D-0531. Preparation of this paper was supported by the Air Force Human Resources Laboratory, Project 744-043. Views expressed in this paper are the author's and do not necessarily reflect the views of the U.S. Army Research Institute or the Air Force Human Resources Laboratory.

scores and predictor data shed light on what each criterion is actually measuring. The purpose of this paper, then, is to examine between-criterion measure and predictor-criterion relationships and to interpret these with the intention of learning more about what scores on the criterion measures really mean.

METHOD

Description of the Performance Measures

A complete description of performance criterion development work can be obtained from other Project A reports. This work included developing the following measures: (1) Army-wide rating scales relevant for evaluating soldiers in any first-tour Army job (Borman, Motowidlo, Rose, & Hanser, 1984; Borman & Rose, 1986); (2) job-specific rating scales (Toquam, McHenry, Corpe, Rose, Lammlein, Kemery, Borman, Mendel, & Bosshardt, 1986); and (3) hands-on proficiency measures and job knowledge tests (Campbell, Campbell, Rumsey, & Edwards, 1986). The Army-wide scales were developed using behaviorally-anchored rating scale methodology (Smith & Kendall, 1963), and focus on performance dimensions relevant to any MOS (e.g., following rules, regulations, and orders; maintaining equipment). The job-specific scales were developed in the same manner; they focus on performance areas more narrowly relevant to a particular job (e.g., loading cargo and transporting personnel-motor transport operator). Finally, hands-on task proficiency measures tap skills in actually completing important tasks relevant to a job, and the job knowledge measures contain paper-and-pencil, multiple choice items assessing knowledge about how to perform the same important tasks.

Administration Procedures

Subjects in the research were 5021 first-term soldiers in nine different Army jobs. Table 1 contains a brief description of the sample.

The rating scales were administered to groups of 15 or fewer peers or supervisors of the target ratees after they were trained using a combination error and accuracy training approach (e.g., Pulakos, 1984). On average, 1.90 supervisor raters and 3.26 peer raters per ratee provided these performance evaluations on the Army-wide and job-specific behavior based rating scales.

Hands-on task proficiency was assessed by administering to each soldier in the sample 15 individual work samples representing 15 of the most important tasks for that job. Experienced job incumbents or supervisors were trained as hands-on scorers, and used a relatively objective checklist to evaluate each soldier on each work sample task associated with that job (Campbell, Campbell, Rumsey, & Edwards, 1986). Job knowledge tests, one for each job, were administered to groups of 15-20 soldiers.

In addition, a specially-developed temperament survey (Hough, Barge, & Kamp, 1985) was administered to all soldiers in the sample. Finally, AFQT scores were available on a data file for a large percentage of the sample.

Data Analyses

For the rating measures, factor analyses were conducted to reduce the number of rating variables to consider. In the case of the Army-wide scales, three varimax-rotated factors were obtained and labeled: (1) Technical Skill, Effort, and Leadership; (2) Discipline; and (3) Military Bearing. We formed unit-weighted composites of the ratings for dimensions loading on each of the factors. Factor analyses of the job-specific ratings yielded results that were difficult to interpret. Accordingly, for

Table 1

Sample Sizes by Job

Job or MOS	N
Infantryman	679
Cannon Crewmember	638
Tank Crewmember	490
Radio Teletype Operator	349
Light Wheel Vehicle Mechanic	597
Motor Transport Operator	646
Administrative Specialist	460
Medical Specialist	481
Military Police	681
	5021

each job, a unit-weighted composite of ratings on all the job dimensions was derived. Likewise, a single hands-on test score was formed by computing the percent of test steps completed correctly and a total percent items correct score was computed for the job knowledge tests. The temperament scales were also factor analyzed, resulting in three summary dimensions: (1) Surgency; (2) Socialization; and (3) Emotional Stability. Unit-weighted composites were derived the same way as they were for the Army-wide rating scales.

Interrater reliability coefficients were computed both within rating source (e.g., peers) and across the peer and supervisor sources. Coefficient alpha reliabilities were derived for the hands-on and job knowledge tests.

Ratings and the objective criterion measures were intercorrelated to evaluate relationships between different methods of assessing performance and to help interpret the meaning of scores on the various performance measures. Also, predictor data on cognitive and non-cognitive scales were available for members of the sample, and correlations between selected predictor scale scores and the different performance criteria also helped to interpret the meaning of performance scores.

RESULTS

Reliability Estimates for the Criterion Measures

For the ratings, interrater reliabilities (intraclass correlations) within rating source are in the mid 40s for peers and approximately .50 for supervisors. Intraclasses for peer and supervisor ratings pooled across sources are .55 - .60. For purposes of the correlational analyses conducted here, peer and supervisor ratings were pooled. Internal, coefficient alpha reliabilities for the hands-on, task proficiency tests range from the 60s to the 80s, and the same kind of reliabilities for the job knowledge tests are in the 80s and 90s.

Relationships Between Criterion Measures

Table 2 presents correlations between the different criterion measures. The correlation between the two relatively objective criteria, hands-on test performance and job knowledge test scores, is .36, whereas relationships between the ratings and the objective criteria are uniformly lower (e.g., .13 and .21 between the composite overall effectiveness rating and, respectively, hands-on and job knowledge test scores). Some construct validity for the rating category composites is derived from the fact that Technical Skill, Effort, and Leadership correlates higher with the objective, maximum performance measures than do the other two categories that conceptually have little relation to the technically-oriented skill and knowledge elements of the objective criteria. However, these correlations only reach .25 and .16 between the Technical Skill, Effort, and Leadership rating composite and, respectively, job knowledge and hands-on test performance.

Relationships Between Predictor Measures and Criteria

Table 2 also reports correlations between temperament and ability predictors and each of the criteria. The AFQT total score, a measure of general cognitive ability, correlates highest with job knowledge test scores. This predictor correlates low positive with the ratings of the individual categories; the highest relationship is with Technical Skill, Effort, and Leadership (r = .14).

On the other hand, the personality predictors are related more highly to the ratings, mostly in the mid-20s for the Surgency and Socialization

Table 2

Correlations Between Predictor and Criterion Measures

(N = 4500 - 5000)

Criteria	1	2	3	4	5	6	7	8	9	10	11
1. Technical Skill, Effort, and Leadership	--										
2. Discipline	73	--									
3. Military Bearing	60	51	--								
4. Overall Effectiveness	86	74	63	--							
5. Overall Job Performance	77	56	49	72	--						
6. Job Knowledge	25	18	02	21	22	--					
7. Task Proficiency	16	06	02	13	17	36	--				
Predictors											
8. AFQT	14	10	-05	11	10	42	10	--			
9. Surgency	28	15	25	26	20	09	03	13	--		
10. Socialization	25	31	24	26	15	11	-04	08	63	--	
11. Emotional Stability	16	12	15	16	14	12	03	16	57	45	--

composites. Correlations between personality variables and the objective criteria are much lower.

DISCUSSION

The pattern of correlations between predictors and criteria provides more information about what these various criteria are measuring. The job knowledge criterion is likely tapping elements of maximum performance, the "can-do" component of effectiveness. A comparatively high correlation with the AFQT predictor further suggests that the job knowledge criterion is reflecting in part a narrower cognitive learning ability aspect of performance.

Moderate sized correlations between the temperament factors and ratings, along with lower such relationships for the objective criteria, suggest that the ratings might be measuring more the motivation-related, effort and hard work components of performance. The highest correlations between individual temperament scales within these composites and overall effectiveness are with work orientation, conscientiousness, and nondelinquency. This further suggests that ratings are tapping the "will-do," try-hard "good citizen" elements of work performance. Referring to the Performance = Ability x Motivation formulation, the job knowledge test is likely measuring the former and ratings the latter. It should be noted that these results are very similar for peer and supervisor ratings taken separately.

It is not clear from Table 2 data what the hands-on task proficiency tests are measuring. Correlations between scores on these tests and all other variables are quite low, with the exception of the .36 correlation with job knowledge test scores. One possible reason for this finding is that analyses were conducted across all nine of the jobs. Different difficulty levels of the proficiency tests for different jobs could artifactually reduce these across-job correlations. The same possibility holds for the job knowledge tests. These possibilities will be explored.

Overall, the different measures of job performance employed here show little -- .vergence across methods. This could be interpreted as a troublesome finding, with error of measurement reducing the between-method relationships to rather low levels. However, data and arguments presented above suggest that the various methods are likely tapping largely <u>different</u> elements of performance. Each method may in fact be measuring its own criterion domain with considerable validity. Campbell, Dunnette, Lawler, and Weick (1970) and Borman (1974) argued that ratings from members of different organizational levels might not agree very closely and yet each source could be providing valid depictions of ratee performance. Extending this argument to multiple methods of measuring performance, lack of convergence across methods may be due in part to the different methods' focus on different aspect of performance.

It should also be noted that confirmatory factor analyt.c work is proceeding in Project A to form criterion constructs that depict the structure of the latent and observed variables using the measures described above as well as additional criterion measures (Campbell, 1986; Wise, Campbell, & Hanser, 1986). This important work is resulting in summary performance constructs that can be used to efficiently and effectively examine predictor-criterion links in the Project A data. Also, path analysis is being employed to examine further the relationships between different criterion constructs and between cognitive and temperament predictor factors and criterion measures (White, Borman, Hough, & Hoffman, 1986). In sum, patterns of correlations between criterion measures and between various predictors and criteria in the present research are providing evidence related to what the different criterion methods are actually measuring.

REFERENCES

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance, 12*, 105-124.

Borman, W. C., Motowidlo, S. J., Rose, S. R., & Hanser, L. M. (1984). Development of a model of soldier effectiveness. In N. K. Eaton, M. H. Goer, J. H. Harris, & L. M. Zook (Eds.), *Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year.* (ARI Research Note 85-4).

Borman, W. C., & Rose, S. R. (1986). Chapter 2: Development of the Army-wide rating scales and task dimensions. In E. D. Pulakos, & W. C. Borman (Eds.), *Development and field test of Army-wide rating scales and the rater orientation and training program.* (ARI Technical Report, in press).

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). *Development and field test of Project A task-based MOS-specific criterion measures.* (ARI Technical Report, in press).

Campbell, J. P. (1986). *When the textbook goes operational.* Invited address to 94th Annual Convention of the American Psychological Association.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill.

Hough, L. M., Barge, B. N., & Kamp, J. D. (1985). Non-cognitive measures: Pilot testing. In N. G. Peterson (Ed.), *Development and field test of the trial battery for Project A.* (ARI Technical Report).

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149-155.

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel R., & Bosshardt, M. J. (1986). *Development and field test of behaviorally-anchored rating scales for nine MOS.* (ARI Technical Report, in press).

White, L. A., Borman, W. C., Hough, L. M., & Hoffman, R. G. (1986). A path analytic model of job performance ratings. In H. R. Hirsh (Chair) *Causal models of job performance.* Symposium presented at 94th Annual Convention of the American Psychological Association.

Wise, L., Campbell, J. P., & Hanser, L. M. (1986). A latent structure model of job performance factors. In J. P. Campbell (Chair) *Multiple criteria and multiple jobs: Will one model fit all?* Symposium presented at 94th Annual Convention of the American Psychological Association.

STANDARD SETTING PROCEDURES:
ARMY ENLISTMENT STANDARDS AND JOB PERFORMANCE

Jane M. Arabian and Lawrence M. Hanser
U.S. Army Research Institute

Presented on symposium,
"Setting Standards in Performance Measurement"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Standard Setting Procedures: Army Enlistment Standards and Job Performance

Jane M. Arabian[1] and Lawrence M. Hanser
U.S. Army Research Institute
Alexandria, Virginia

Introduction

The Army Research Institute is currently engaged in a large-scale, multi-year research project to improve the Army selection and classification system (Project A, "Improving the Selection, Classification and Utilization of Army Enlisted Personnel") and, thereby, increase the overall effectiveness of the force. The research is aimed at developing comprehensive selection and classification procedures to predict validly performance in Army training and occupational specialties.

A number of performance measures, including measures of training success, service-wide performance, and MOS-specific hands-on performance, were developed. The Army's rationale for developing multiple measures of job performance is based upon the knowledge that a soldier's job is multifaceted (i.e., many different kinds of tasks are involved) and there are multiple aspects to job performance (e.g., initiative, obedience, etc.). Therefore, in order to obtain information about the domain of job performance behaviors, the Army's research project has developed different kinds of tests to assess these different aspects of job performance. Composite

[1] The opinions, views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, expressed or implied, of the U.S. Army Research Institute for the Behavioral and Social Sciences or the Department of Defense or the United States Government.

scores, based on constructs derived from the performance measures, will be used as indices of job performance.

Preliminary analyses of field test data and other research lead to the expectation that data from the concurrent validation phase of the project will result in positive correlations between predictor and criterion measures of performance. This information, however, will not lead directly to the setting of enlistment standards. While it is possible to use cost trade-off models for selection and classification systems once a performance objective is determined, such models cannot identify the required performance objective. Some other method is needed to define performance requirements or standards before reasonable enlistment standards can be established.

To determine whether existing methods could be used to set job performance standards in the Army, the literature review in Appendix A was conducted. It is in the form of an annotated bibliography summarizing the content of each reference. While the bibliography is not intended to be exhaustive, it is representative of the published literature. Table 1 presents a listing of the bibliographic references and indicates the broad categories that reflect the content of each reference.

Overview of Standard Setting Procedures: General Issues

The majority of the references present definitions of standard setting and descriptions of various procedures that have been developed. The article by Glass (1978) describes a variety of standard setting methodologies. In addition, there are individual references for methodologies developed by Angoff (1971), Jaeger (1976, 1982, 1984) and Nedelsky (1954).

Table 1
Content Description of the
References

Reference	Issues/ Critique	Definitions	Description of Procedures	Comparison of Procedures	Education/ Certification	Job Performance Standards	Application (Specific)	"Utility"/ Decision Models
ndrew, et al (1967)				X	(X)		X	
ngoff, (1971)			X					
Block, (1978)	X							
Buck, (1977)	X	X	X					
Buck, (1975)		X						
Burton, (1978)		X	X					
Chuang, et al (1981)			X					X
DuBois, et al (1954)		X						
Eastman, (1981)						X	X	
Glass, (1978)	X	X	X					
Hambleton, (1978)		X			X			
Hambleton, et al (1980)		X	X		X			
Hambleton, et al (1979)			X					
Hofstee, (1983)	X		X					
Jaeger, (1982)			X	X	X			
Jaeger, (1976)	X	X						
Jaeger, et al (1984)				X	X		X	
Koffler, (1980)				X	(X)		X	
van der Linden, (1980)								X
Linn, (1978)	X							

Table 1
(Continued)

Reference	Issues/ Critique	Definitions	Description of Procedures	Comparison of Procedures	Education/ Certification	Job Performance Standards	Application (Specific)	"Utility"/ Decision Models
vingston, et al (1983)				X	X		X	
cPherson, (1981)			X			X	X	
aslow, (1983)	X					X		
essick, (1975)	X				X			
edelsky, (1954)			X					
ngrio, (1984)				X	X		X	
opham, (1978)	X	X						
Reid, (1984)			X	X				
Scriven, (1978)	X					X		
Shepard, (1984)	X		X					
Shepard, (1983)	X		X					X
Shepard, (1976)	X	X						
Shikiar, et al (1985)	X					X		

The Poggio (1984) article provides a good comparison of various procedures used in Kansas to set standards on educational competency tests for reading and math. A large group of references deal with applications of standard setting procedures. By far, the bulk of the applications deal with setting educational standards (e.g., minimally acceptable levels of reading and math knowledge for high school graduates) and professional certification (e.g., minimal levels of knowledge that a grammar school teacher must possess to be certified or licensed to teach in a given state). There is clearly very little empirical, applied research dealing specifically with the determination of job performance standards for selection and classification purposes.

Standard setting for certification vs. selection and classification. A basic difference between standards for competence (mastery vs. nonmastery) or certification (CC) and for selection and classification (SC) is that the former essentially entail only one judgement. A CC standard is used to indicate that an individual meets the qualifications to be considered minimally competent.

Standard setting for SC purposes requires that two judgements be made. The first is a judgement of minimal competence or acceptable performance derived through measurement of job performance with job incumbents. In other words, a standard needs to be set on the criterion measure(s). A second cut-off score (standard) needs to be set on some predictor measure such that individuals who meet the standard on the predictor will be likely to meet, at some later point, the standard for acceptable job performance. Actually, two sets of these types of judgements may be needed: One for selection and a second for classification. The first set of judgements, for selection, may require different considerations than the set of judgements for job classification. For example, selection standards may be based on concerns regarding supply and demand, trainability in a general sense and attrition, while classification standards would be based upon considerations of actual job performance. The Army model is based on this sort of multiple judgement approach.

Procedures: judgement and validity. Regardless of the context for setting standards, it should be understood that the application of any standard setting procedure requires judgement. While the judgements will be value-laden, they are not therefore arbitrary in the sense of being based solely on whimsey (cf. Hofstee, 1983). The judgemental nature of standard setting has been a focus of debate in the literature (e.g., Glass, 1978; Hambleton, 1978). However, rather than dismiss all standard setting procedures because they require judgements we need a more constructive approach. It seems reasonable to accept the fact that judgements are the basis of standard setting procedures and then examine the validity and impact of the resultant cut-off scores.

Unfortunately, the literature offers very little guidance for selecting one procedure over another based on considerations of validity. Indeed, it cannot be said that a test performance standard derived from any one procedure is intrinsically valid because of the particular procedure employed. Andrew and Hecht (1976) found that different groups of judges arrived at similar standards using any one procedure, but different standards were obtained when two different procedures were used by the same groups of judges. According to Poggio's (1984) research, different procedures will consistently yield higher or lower standards. In fact, the validity of a cut-off score (i.e., the ability of the cut-off score to discriminate between minimally acceptable and unacceptable individuals) is likely to depend not only on the procedure used to set the cut-off but also on the content of the instrument(s) used to assess performance.

Norm-referenced vs criterion-referenced tests. It stands to reason that if a test does not suitably measure what it purports to measure then any standard or cut-off score based on that test will not be valid. This holds for tests developed within either a norm-referenced test (NRT) or criterion-referenced test (CRT) framework. Although a detailed description of NRT and CRT development procedures will not be presented here, several references in the bibliography (see Table 1, "Definitions") discuss the methods in more detail.

One point that bears emphasizing is the different goals of the tests. Basically, a NRT is designed to optimize discriminability between all individuals administered the test. A CRT is designed to maximize discriminability around the cut-off point for proficiency, and, technically, is composed only of items necessary for identifying proficiency in the content domain being tested. In terms of test item difficulty, NRTs tend to contain a range of item difficulties; CRT items, on the other hand, are viewed as homogeneous.

Despite different test development strategies, both types of tests will yield a distribution of response scores (e.g., number or percent correct). However, with a NRT one expects to obtain a more normal distribution of scores while with a CRT one expects a somewhat skewed and peaked response distribution on, e.g., an end of course exam. It must be stressed that since ability is a continuous, not all-or-none, variable, test scores will always reflect a variety of abilities. Variations in scores are not simply measurement error on either a CRT or NRT. Therefore, scores on either type of test may be given a norm-referenced interpretation. Just as pass/fail standards are set on both NRT and CRT, one can discuss an individual's score in relation to all other scores from the exam with either a CRT or NRT.

With respect to the selection of a standard setting procedure, any procedure can be applied to either type (NRT or CRT) of test. However, the literature does imply that once a cut-off score or standard is set on a CRT, one may denote individuals whose scores fall above the standard as "masters" of the domain covered by the test. Individuals whose scores fall below the cut-off are designated as "non-masters" of the subject matter. In fact, one specific purpose of a standard on a CRT is to identify an individual as either a master or non-master of a particular skill domain (though not, necessarily, as more or less masterful than another individual). Appropriate labeling of individuals scoring above or below the cut-off point on an NRT is less clear. This may be attributed to the fact that since NRT items are sampled statistically (i.e., randomly from a pool) rather than on strict content domain grounds, it is less clear what an individual would be a master of, except the items on the test. The inference from the test items to the domain of skill is weaker for a NRT than for a CRT.

The report by Buck (1977) provides a good discussion of concerns and standard setting procedures for NRTs and CRTs. By way of summary, Buck states that "a test is not inherently norm-referenced or criterion-referenced. It is the manner in which a test is developed and interpreted that determines whether it is to be classified as norm- or criterion-referenced. It is conceivable that a test could be either norm- or criterion-referenced or both depending on the way in which it is developed, used and interpreted [p. 15]". Indeed, the Army's Project A measures encompass both NRT and CRT aspects. Due to the scope and purpose of the project, the development of the criterion (job) measures was based upon careful, comprehensive identification of the criterion domain followed by non-random sampling of tasks within the domain and construction of test items in such a way as to optimize discrimination of ability levels among individuals[1].

Modes of measurement. By and large, the published literature on standard setting deals with paper and pencil, multiple choice (recognition) tests. Applications of standard setting procedures to, for example, hands-on, rating, or interview assessment procedures are not represented in the literature (cf. Shikiar, et al 1985). This is not to say, however, that the existing procedures or the principles they embody cannot be made to accommodate different testing modes. Standard setting procedures can also be augmented to encompass the practice of using multiple tests and multiple test modes for criterion (job) performance measurement. Alterna-

[1] For a more detailed description of the criterion measures see: Campbell, C.H., Campbell, R.C., Rumsey, M.G., and Edwards, D.C. (1985). Development and field test of task-based MOS-specific criterion measures. Alexandria, VA: US Army Research Institute, in press.

tive approaches are described in Buck (1977).

Selecting and Applying Standard Setting Procedures: General Considerations.

Although the literature does not offer specific guidance on selecting one procedure over another for different situations, general recommendations or areas for consideration can be identified. Specific areas will be addressed in the following paragraphs: Acceptance of standards and modifying standards.

Acceptance of standards. One important consideration is the selection of the judges (standard setters). If representation of the end-users is included in the standard setting process, the likelihood that the resultant standards will reflect the interests, concerns, and needs of the users is increased. In the context of selection and classification in the military, it would be prudent to include individuals from the personnel, training, policy and field communities on standard setting panels.. Involving several judges or groups of judges in the standard setting process may help to promote confidence in and acceptance of the standards. When independent groups of judges employ the same standard setting procedure and arrive at similar standards (cf. Andrew and Hecht, 1976), confidence in the standard will be increased.

Another consideration for increasing acceptance of standards is related to both the judges involved and the procedure selected. A procedure that seems convoluted to the judges or asks them to make decisions they do not feel qualified or knowledgeable enough to make is unsatisfactory. However, the very same procedure presented to a different group of judges may meet with a more satisfactory response. Any procedure that causes judges to feel uneasy is not likely to result in a cut-off score that users will feel confident in implementing; the validity of the standard will be called into question. This is not to say that judges should be selected to "fit" the procedure. Rather, it is recommended that a procedure should be selected to "fit" the standard setters. Every effort should be made to ensure that that judges find the procedure credible and easy to apply.

Several sources have suggested providing judges with normative data so that their expectations of performance will not be unreasonable (e.g., Livingston and Zeiky, 1983; Shepard, 1976). Incorporation of normative data is likely to result in similar standards across judges which, in turn, is likely to improve confidence in the selected standard. Further, it should be noted that Jaeger et al. (1984) have found that iterative applications of a given standard setting procedure result in reduced variability across judges. Iterative applications, however, did not affect the mean recommended standard. The reduction in variability, i.e., better agreement among judges, is also likely to increase the confidence of the judges in the resultant standards.

Modifying standards. The preceeding discussion has concentrated on ways to maximize confidence in standards derived by any given procedure. It is important to ensure not only that the judges themselves are confident with the standard but also that the end-users and individuals directly affected by the use of the standards accept the results. Therefore, in addition to the above considerations, institutional requirements and values must also be taken into account. No matter how confident judges may be in their decisions, if the standards appear too high or too low from an institutional perspective the standards will not be acceptable.

It was stated earlier that some procedures consistently yield higher standards than others. This means that some procedures will result in relatively high standards that are likely to produce false negative decisions, i.e., individuals will be classified as not minimally acceptable when, in fact, they would have been able to perform at a level acceptable to the organization. Conversely, procedures resulting in lower standards are more likely to produce some amount of false positive decisions. At an organizational level, then, consideration may be given as to whether false negative or false positive decisions are more serious or costly to the organization.

Decision theoretic and utility analyses can serve as a tool to "fine tune" a standard set by panels of judges. Decision theory is not a standard setting procedure; it is a technique for reducing the effects of measurement and sampling error (van der Linden, 1980). The goal of utility analysis is "to match the test dichotomy to the criterion dichotomy to ensure that the smallest number of classification errors will be made" (Shepard, 1983). One form of utility analysis is to determine the relative cost of one kind of error (false negative) against the cost of another kind of error (false positive). This may be a difficult, complex approach to apply especially with respect to deciding which cost factors should be used (e.g., cost of training, equipment loss, dollar value of performance).

Eaton et al. (1985)[1], have presented utility estimation techniques developed to be easier to apply in situations where "managers are more accustomed to considering the relative productivity of employees or crews than the costs of producing given levels of output ...[or]... where employees operate very complex, expensive equipment and/or are focal to the productivity of a costly system [p. 29]". The strategies presented by Eaton et al. consider changes in the number and performance level of system units for increased aggregate performance. As noted by the authors, these techniques still do not provide for easy linkage of performance quality to a single quantitative scale. The "linkage" maybe require complex judgements regarding the utility equivalence of different performance levels for different situations or groups of individuals.

Within Project A, attempts are being made to scale the value of different levels of performance. Utility scaling workshops will be conducted with military personnel. Their task will be to scale different performance levels of various Army occupations using the 50th percentile performance of the infantryman occupational speciality (11B) as a baseline. It is conceivable that the resultant scale value for the utility of individuals performing at the 90th percentile in some occupations will be lower than the scaled utility of the 50th percentile 11B.

Once the utility of performance levels is scaled onto a single dimension, information obtained from the scale may be used to modify test performance and/or entrance standards for different occupations in order to

[1] Eaton, N.K., Wing, H. and Mitchell, K. (1985). Alternative methods of estimating the dollar value of performance. Personnel Psychology, 38, 27-40.

optimize selection and classification decisions. If, as expected, a given performance level (standard) does not have the same utility across occupational specialties, then the classification standards for each occupation may be modified in order to optimize the overall utility of the total enlisted force.

Conclusions

This review has started from the premise that tests on which standard setting procedures are applied have already been determined to be psychometrically sound. This is to say that the test must be valid, reliable, and follow the guidelines of the American Psychological Association[1] for test development practices. Once a test has been appropriately developed, a test performance standard (cut score) can be set.

Each standard setting procedure should be applied judiciously and care should be taken so that the mathematics involved in some of the procedures do not create a false sense of rigor. Every standard setting procedure is based on judgement. It is the responsibility of the developers, users, and overseers of the standard setting process to ensure that the judgements are sound, appropriate to and supportive of the goals and values of the organization or community served by the standards (cf. Hofstee, 1983). Further, it is incumbent on the responsible parties to evaluate and reevaluate the standards in terms of the impact the standards have on the organization. The basic objective of any standard is to help attain the

[1] American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: Author.
American Psychological Association, Division of Industrial-Organization Psychology. (1980). Principles for the validation and use of personnel selection procedures. (Second edition) Berkeley, CA: Author.

critical goals and requirements of a given institution. Jobs, for example, change as do the needs of the organization. Accordingly, measures of job performance and standards based on those measures must be regularly appraised and modified as needed to ensure that the values of the organization are being met.

The concluding point of this review is that there cannot be one and only one correct standard. The notion that one correct standard can be determined for a given situation is logically inconsistent with the fact that performance or ability exists as a continuous variable. Any standard, no matter how it is derived, imposes an artifical dichotomy (e.g., pass vs fail, master vs non-master, etc). This not to suggest that standards should be eliminated or avoided. Standards do serve as useful tools in the selection and classification processes. Rather, the standard setting process cannot end with the determination of a particular standard. There is a need to continue evaluating the standard to ensure that the number and cost of the inevitable decision errors produced by that standard are minimized.

APPENDIX A

ANNOTATED BIBLIOGRAPHY

ANNOTATED BIBLIOGRAPHY

Andrew, B.J. and Hecht, J.T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-5C.

Abstract

Two standard setting procedures were employed by two groups of judges to set pass-fail levels for comparable samples of a nationally administered examination. These procedures were both designed to set standards in relation to the minimally qualified examinee. The study was undertaken to determine whether similar standards would be set for the same examination content when determined by different groups of judges, and whether the two procedures employed would result in similar standards for comparable samples of test content. In addition, the extent to which group consensus judgments might differ from individual judgments was also investigated. The results suggest that different groups of judges do set similar examination standards when using the same procedure, and that the average of individual judgments does not differ significantly from group consensus judgments. Significant differences were found, however, between the standards set by the two procedures employed. This finding was observed for both groups. The nature of these differences is described, and their implications for setting examination standards are discussed. (Author)

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement (pp. 508-600). Washington, DC: American Council on Education.

- Pg. 514-515: description of procedure:

Systematic procedure for deciding on minimum raw scores for passing and honors - think of "minimally acceptable person"

- go through test item by item;
- could such a person answer correctly the item under consideration: correct score 1
  incorrect score 0
- sum of scores = raw score [cut off] of minimally acceptable person
- have some number of independent judges decide by consensus without actually administering the tests
- results could later be compared with numbers and percent of examinees who actually earned the passing grades [validity studies: verify appropriateness of the initial cutting scores or correct them if necessary]
- or ask each judge to state the probability that the minimum acceptable person would answer each item correctly; the sum of the probabilities would represent the minimum acceptable score

- Pg. 531...suggestion of applying cut score procedure to Army.

Block, J.H. (1978). Standards and criteria: A response. Journal of Educational Measurement, 15, 291-295.

- Responds to Glass (1978) paper

- Argues that standard-setting techniques are not as arbitrary as Glass suggests

- Suggests developing new and better technique. ....promote broad-based humanistic procedure.

Buck, L.S. (1977). Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures (Technical Memorandum 77-4). Washington, DC: Personnel Research and Development Center United States Civil Service Commission.

- Presents a review and summary of methods for establishing cut-scores

- Discusses issues as they apply to norm and criterion reference tests

- Pg. 9: summary of issues and models of test fairness
- Pg. 13-15: summary - cutting scores for NRTs
- Pg. 20-21: summary - cutting scores for CRTs
  (10 "models"... methodology different ... but no empirical or theoretical basis for selecting one over another)

- Pg. 13: "The test developer must realize that the process of setting a cutting score cannot be totally analytic, as it is impossible to assume a purely objective attitude".

Buck, L.S. (1975). Use of criterion-referenced tests in personnel selection: A summary status report (Technical Memorandum 75-6). Washington, DC: Personnel Research and Development Center United States Civil Service Commission.

- Discusses validity of CRTs and measures of reliability (actually does little more than reference papers dealing with the topics)

- See pg. 21 ... measures of reliability.
  pg. 23 ... validity
  pg. 26 note ... reliability/correlation estimate

- Provides 28-page annotated bibliography

Burton, N.W. (1978). Societal standards. Journal of Educational Measurement, 15, 263-272.

- Pg. 264 - definition of criterion - differences in emphasis:
  1) criterion [variable] - trait to be measured (traditional definition)
  2) - specification of minimum levels of performance (Glaser and CRTs)

- Pg. 266 - 3 types of methods: 1) standards based on theories (learning hierarchies)
  2) standards based on expert consensus
  3) standards based on practical necessities (minimal competencies for real-life)

Chuang, D.T., Chen, J.J., & Novick, M.R. (1981). Theory and practice for the use of cut-scores for personnel decisions. Journal of Education Statistics, 6, 129-152.

- Mathematical model

- Optimize utility ... final cut-score set by utility function...
  ... assumes some cut score has already been determined...

- Does not specify how cut scores are set

DuBois, P.H., Teel, K.S., & Petersen, R.L. (1954). On the validity of proficiency tests. Educational and Psychological Measurement, 14, 605-616.

- A proficiency test is considered valid if it discriminates between the proficient and non-proficient in a given skill... "while an aptitude test may have an indefinite number of validities, depending on the criteria which it predicts with varying degrees of success, a valid proficiency test must measure what it purports to measure. No other concept of validity is applicable, the only variation possible is in the method used in arriving at the estimate of validity" (pg. 605)

- Coverage and discrimination power are independent dimensions of a proficiency test

- Difficulty analysis - item has p-value of .50 (passed by 50% of the group)... its SD is at maximum and makes maximum numbers of discriminations... A range of item difficulties, from very easy to very hard with a mean at about .50 is optimal for differentiating within a given population

- Types of validity:
  - Validity by "Direct Judgment" -- uses SMEs
  - Work sample validity -- correlation with work sample that is representative and meaningfully measures the skill
  - Class validity -- e.g. high vs. medium vs. low proficiency
  - Curricular validity - untrained vs. trained

Eastman, R.F. (1981). Supervisor ratings as criteria for Skill Qualification Tests. In S.F. Bolin (Chair), Panel on skill qualification testing: An evolving system (pp. 1356-1366). Arlington, VA: 23rd Annual Conference of the Military Testing Association. [DTIC, ADP 001400]

- Correlation between supervisor ratings of overall job performance and SQT scores for 67N (r=.74)

- Used ratings as criteria to determine optimum cut score for performers vs. non-performers

  ratings:
  - competent/not competent/don't know
  - cross out name if you don't supervise the individual
  - use a "+" sign to indicate one of the best soldiers and a "-" sign to indicate one of the poorest soldiers

- Each soldier rated by 3-6 supervisors

- Plot (tabulate) distribution of performers (rating ≥ 3) by SQT score

- Findings: SQT cut-score could be lowered and be more consistent with perceptions of supervisory personnel; it may be better to have supervisor rank-order soldiers instead of having to designate a soldier as a non-performer.

Glass, G.V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-262.

- Pg.243: Criterion: the definition has been corrupted; originally referred to a criterion-referenced test ... meaning a scale of behavior linked to a test scale ... now criterion taken to be synonymous with "standard" or "cut score"

- Glass expresses strong concern regarding the arbitrary nature of setting a standard and the notions of a standard

- Six classes of methods for establishing cut-offs:

1. Performance of others - reference parameters of existing population of examinees - e.g., median score, 50th percentile ... essentially normative... not "behaviorally informative"; criterion-reference test theorists would find this approach to be an inappropriate method.
2. "Counting Backwards from 100%" - given the nature of criterion-references test (objectives)...would expect perfect scores...but allowances must be made for, e.g., measurement error and clerical mistakes...but how much??...the method is highly judgmental and too vague
3. Bootstrapping on other criterion scores - use other determinations of competence to select a group then match the group against the score distribution of some other test...problems 1) the 2 tests must be correlated, but it will never be perfect, therefore, you will make, e.g., false positive and false negative decisions, and you still have the problem of 2) how was the standard set on the first test...circularity problem
4. Judging minimal competence - study a test and determine the required score for a minimally competent individual (cf. Nedelsky; Ebel)...see pg. 246 for Nedelsky method; pg 247 for Ebel method; pg. 248 for Angoff method; problems: 1) consistency and reliability of judges; 2) logical psychological status of concept of minimal competence.
5. Decision theoretic approaches - cutoff on an external criterion assumed as a "given"...vary score on the criterion-referenced test to say, minimize false negatives...approach simply postpones decision regarding the setting of a cut off... still "arbitrary"
6. "Operations Research" Methods - based on OR approach of maximizing a valued commodity by finding an optimum point on a mathematical curve or graph--- must have a non-monotonic curve...could have composite with a second valued outcome; but then have the problem of how to weight the composite...or look for the point of diminishing returns...(no further gain) how do you decide non-arbitrarily

- Pg. 258... standard-setting procedure may involve more precision than the test itself has... no matter what procedure used, there is still the element of the arbitrary

- Glass favors a comparative approach, e.g., improvement (change in performance) but you still have the questions re: how much change is good/-sufficient...how much loss before action should be taken... same problem as with criterion score, but he claims one has still gained clarity and consensus even if all problems were not solved.

Hambleton, R.K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

- Validity of cut score depends on how accurately it separates examinees into mastery states...usually the criterion is some external measure of performance or instructed vs. non-instructed groups

- Methods are based on the consideration of item content, educational consequences, psychological and financial costs, performance of others, errors due to guessing and item sampling...all arbitrary

- Against Glass' (1978) recommendation of using change scores.

Hambleton, R.K. & Eignór, D.R. (1980). Competency test development, validation, and standard setting. In R.M. Jaeger & C.K. Tittle (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 367-396). Berkeley, CA: McCutchan Publishing Corporation.

- Focuses on making competency judgment for individuals - not groups (e.g....program evaluation)

- "A minimum competency test is designed to determine whether an examinee has reached a prespecified level of performance necessary to each competency being measured"..."standard" [or "cutoff score" or "minimal proficiency level"] is a point on a test score scale which is used to separate examinees into two categories"...master/nonmaster - a standard is set for each competency measured by a test..."competency tests are a special type of criterion-referenced test" - requires "information about levels of individual performance relative to well-defined content domains (referred to as "domain specifications)"

- 4 important topics: 1) improved guidelines for preparing domain specifications
  2) guidelines for evaluating competency tests and test manuals
  3) research on the relationship among test length, test score reliability, and test score validity
  4) consideration of issues and methods for determining standards, as well as guidelines for implementing each method

- Goes through 12-step model for developing and validating competency tests

- Pg. 377: test length formula for criterion reference tests, (vs. Spearman-Brown for norm-referenced tests)

- Continuum vs. state models (all-or-none): in the latter, test true-score performance is viewed as all-or-none, true-score standard is set at 100%, after consideration of measurement error the observed-score standard is set at a value less than 100%...use normative information as an aid in making decisions (in case experience of judges may have been with unusual students)

*See pg. 383-384...different models' use of (need for) utility values
pg. 386 - comparison of standard setting procedures
pg. 392 - for empirical methodology and the need for external criterion measures (see refs.)
pg. 3 - latent trait models...feasibility with competency tests??, equating scores from one form of competency test to another.

Hambleton, R.K. & Eignor, D.R. (1979). Issues and methods for standard-setting. In AERA Training Program MAterials, Criterion-referenced test development and validation methods (Unit 6). Unpublished training materials.

- Very similar to Hambleton and Eignor (1980)

- More step-by-step detail (how to do as well as in-depth comparison of methods)

- Reviews different methods (descriptive)

- References Berk (mathematical) methodology and provides algorithms for maximizing correct decisions and minimizing incorrect decisions

- Pg. 47 - describes Livingston method and the use of performance data vs. judgmental data on performance; stresses need for research

- Pg. 51 - to use Black's optimization strategy...need weight for valued outcome criteria to form composites...no specifications for how to do that...further problem: solutions are likely to be situation specific

- Summary (pg. 57)

- If object is to view a test by itself and not in relation to other variables, use Angoff or Nedelsky methods

- If empirical data are available, use Berk or Contrasting Group method

- Pg. 57-59: Hambleton's guidelines...use several groups of judges working together; work through practice examples (with Ebel or Nedelsky method); introduce domain specifications; schedule time to discuss each specification; make sure judges know how the tests will be used;...look at consistency of different groups cut-off score ratings; use performance data to modify cut-off scores; check back to see if objectives are "out of line"; try to compare mastery status of instructed and uninstructed groups of examinees; re-review cut-off scores periodically since priorities change.

Hofstee, W.K.B. (1983). The case for compromise in educational selection and grading. In S.B. Anderson & J.S. Helmick (Eds.), On educational testing (pp. 109-127). San Francisco, CA: Jossey-Bass Publishers.

- Selection practices are political in the sense of promoting certain values at the expense of other values

- Compromise between politics, as referenced above, and scientific fact

- Example of Weighted Lottery procedure for restricted admissions

- Chance of being admitted is a monotonically increasing function of e.g., grade point average - see pg. 113 (illustration)

- Pg. 118 - Fig. 3 - compromise model for establishing cutoff points

- Determine maximum and minimum acceptable % mastery (k) and maximum and minimum acceptable % failures (f); solve f(min) and K(max) and f(max) and K(min); e.g. (0, 70) and (60, 40); locus of admission cut off scores is k+.5f = 70 [k+af=C]

- Actual cutoff is the point of intersection between the model (k+af=c) and the empirical curve.

Jaeger, M. (1982). High school competency test standards and the definition of competence. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY. (ERIC Document 220-478).

- Examines question regarding implicit definition of competence and the inferential chain that links the standard setting process to the decision outcomes of the method for 2 classesof standard setting procedures 1) data-free judgments of items (Angoff, Nedelsky, Ebel) and 2) data-based judgments of items (Jaeger)

- Re: Angoff method: judges must conceptualize the competency construct, i.e., the minimal capabilities required for functioning...not just ability; this would include motivation, initiative, social status, persistence, discipline and other constructs that "frame the competence construct" and then estimate success probabilities on items;method does not prescribe how judges should be selected, minimum number of judges, or qualifications of the judges, given all the variables, claims of (construct) validity of the competency test (or the procedure) are difficult to substantiate

- Re: Nedelsky Method: similar to Angoff re: inferences required for validation...but Nedelsky method requires more specific inferences since each judge must consider each item option and decide whether or not a minimally competent individual would know if it was correct or incorrect

- Re: Jaeger Method: for each item, must decide whether or not every high school graduate should be able to answer the item correctly...judges are given a variety of evidence on past performance of examinees and consequences of a variety of standards; this is less demanding than other procedures (since no conceptualization of minimally competency or no estimation of performance of examinees on the test is required), but requirements for construct validation are the same as for other methods

- Summary: no idea of validity of any procedures used to set standards for high school competency tests; also lack evidence of construct validity of competency tests; this is no indication that any method is any easier to validate than any other method.

Jaeger, R.M. (1976). Measurement consequences of selected standard-setting models. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

- All standard setting is judgmental; the difference between methods is primarily in terms of the proximity of the judgment-determining data to the original performance; "judgmental models" can be viewed as "direct" or "proximal," and "empirical models" as "distal" or "derived"

- Two kinds of inferences "inferences to the performance of individuals on a well-prescribed domain of tasks, and inferences to the performance of individuals on some ultimate criteria that lie outside a sampled domain"

- The first kind of inference has 4 possible sources of error: random error among judges who set standards for domain performance, bias error due to inappropriate sampling of tasks, error due to the description of tasks in a domain, and random error due to an inadequate sample of tasks; all four sources of error threaten the validity of inferences

  - Need to consider consequences of decisions
  - Need a theory of validity to emerge as well as guides to practice

- See pg. 26: matrix of standard setting models and Procedures X Threats to Validity

Jaeger, R.M. & Busch, J.C. (1984). The effects of a delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations. Paper presented at the Joint Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA. [ERIC Document 246 091]

- Procedure:
  - I.
    - a. Have judges take the test/subtest
    - b. Apply Angoff procedure
  - II.
    - a. Give instructions on meaning and interpretation of item difficulty values (practice)
    - b. Allow the opportunity to reconsider their initial estimates (provide data)
      1. Estimate difficulty (from previous administrations of the test by ETS) for each item
      2. Estimate difficulty from sub-population of examinees whose total scores are close to the passing scores established in previous administrations
      3. Produce a cumulative distribution function of sub-test scores
      4. Provide their response sheet from the 1st judgment session, with the recommended standard derived from their judgments; split raters into two groups:

         Silence
         Discussion (controlled discussion; present reasons underlying 1st session judgments)
         Allow judges to modify their passing score determinations

- Greater reduction in variance for "Discussion" group than "Silence" group was found

- Iterative judgments also reduced variability

- Reduction in variability did not appear to have a significant effect on the mean recommended standards

- But, N.B., the study was conducted with a small sample size; precision of estimates of means and variance is questionable, nevertheless, consistency of results is compelling.

Koffler, S.L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

- Used Nedelsky and Contrasting Groups methods to set standards (on New Jersey Minimum Basic Skills tests in reading and math)

- Conclusions: "no substantial agreement or pattern of disagreement between the cut-off scores developed by the Nedelsky and the Contrasting Groups methods" was found

- Since "mastery" is continuous, not dichotomous, no model provides a scientific means for discovering the "true" standard

- Therefore, must carefully analyze data, judgments, and extraneous conditions in any particular situation which may affect the estimates for a procedure and use a variety of procedures

- This paper does not recommend a method for consolidating a variety of different standards from different procedures for a particular test.

van der Linden, W.J. (1980). Decision models for use with criterion-referenced tests. Applied Psychological Measurement, 4, 469-492.

- Bayesian approach

- Decision theory can not be used to set true score cutoffs, but can be used to set observed cutoff score once the true cut-off score has been set

- Decision-theoretic approach to criterion-referenced testing is not a standard-setting technique, but a technique to minimize the consequences of measurement and sampling error (if the true cut-off is 16/20, this approach helps you to choose the observed cut-off: 19/20).

[Also discusses the practicality for use in the military; sample size considerations; potential problems with the interpretability of results]

Linn, R.L. (1978). Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 15, 301-308.

- "A standard often is assumed to have been established and then is used as the starting point to develop techniques for determining whether an examinee should be classified above or below the standard...little attention has been directed to the questions of where the standards come from, who establishes them, and what procedures are used to set them."

- It is desirable to have precision in the definition of the content domain, as with criterion-referenced tests, as well as comparative data from norm-referenced tests

- There is a misleading simplicity to setting standards; therefore, the process should be iterative with different groups of judges.

Livingston, S.A. & Zeiky M. (1983) *A comparative study of standard-setting methods* (Research Report No. 83-38). Princeton, NJ: Educational Testing Service.

Abstract

The borderline-group method and the contrasting-groups method were each compared with Nedelsky's method at four schools and with Angoff's method at another four schools, using tests of basic skills in reading and mathematics. The borderline-group and contrasting-groups methods produced similar results when approximately equal numbers of students were classified as masters and nonmasters. The contrasting-groups passing score was lower than the borderline-group passing score when masters greatly outnumbered nonmasters; higher when nonmasters outnumbered masters. Results involving the Nedelsky and Angoff methods were not consistent across schools. Passing scores tended to be higher at schools where students were more able. (Author)

- Judges were reading and math teachers for grades 6,7,8 and judged reading and math tests, respectively

- For the study, "mastery" was defined as "the ability to perform adequately the reading/mathematical tasks of adult life in modern American society. These tasks were not specified or enumerated...There was no suggestion of a relative standard"; the "borderline" test-taker was defined as is "one whose knowledge or skills measured by the test is on the borderline between sufficient and insufficient".

- Recommendations: 1) with Nedelsky and Angoff methods, consider a modification that allows judges to revise their judgments on the basis of actual student response data from the test; 2) results may have been different if teachers at each school had been required to agree on a precise verbal definition of the standard in behavioral terms before judging their students on the test items..."this step could be the missing link that provides for consistency between standard-setting method based on judgments about students [e.g., contrasting or borderline groups methods] and methods based on judgments about test questions [Angoff or Nedelsky methods]".

MacPherson, D. (1981). Predicting skill qualification test item difficulty from judgments. In S.F. Bolin (Chair), Panel on skill qualification testing: An evolving system (pp. 1383-1390). Arlington, VA: 23rd Annual Conference of the Military Testing Association. (DTIC, ADP 001400)

Abstract

Judgments of item difficulty by small groups of three to six non-commissioned officers were compared with observed item difficulties among soldiers in three military occupational specialties representing infantry, engineer and administrative career fields. Linear correlations between average judgments and observed difficulties were on the order of .50, but the scatter plots were triangular in appearance because objectively easy items were rarely judged to be difficult while objectively difficult items yielded a wide range of judged difficulties. Hence sets of items showing wide and fairly flat distributions of difficulty had been judged to be skewed toward the easy end of the difficulty distribution. These analytic observations suggest that NCOs involved in test construction may be making tests more difficult than they believe, and that NCOs as trainers preparing soldiers for their SQTs may be underestimating the need for training. If the triangular relationship between judged and observed difficulty is confirmed in larger samples of items, then a simple expectancy table method might be used to predict objective test difficulty and training need. (Author)

- N.B. with regard to methods for standard setting that implicitly (or explicitly) require judgments of item difficulty, e.g., Angoff method...supports the position of e.g., Livingston and Zeiky (1983), that actual student response data be used in conjunction with the standard-setting methods.

Maslow, A.P. (1983). Standards in occupational settings. In S.B. Anderson & J.S. Helmick (Eds.), On educational testing (pp. 91-108). San Francisco, CA: Jossey-Bass Publishers.

Research needs:

- Examine the process of human judgment and "practical" questions such as how many judges are needed, etc. as well as the issue of the adequacy of job analysis

- "For assessing competence, a statement of work behaviors is needed that is relevant to the concept of competence"; the behaviors can be "identified by factor or cluster analysis of task questionnaires, by critical incident studies, or observational techniques" and should "include not just what things are done but expectations as to the manner, impact, quality, or consequences of the work"...these are the "core of standards"..."to the extent that these critical behaviors are independent (i.e.,) call for different characteristics, the assessment model must be multidimensional"

- Must examine reliability and validity for all sorts of measures

- With respect to generalizability, should consider the probability that the constructs defining competent performance may be more readily generalized than the specific tests and measures of those constructs.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

- Deals with education, not job performance issues; however, provides a good discussion, scientific but not heavily technical, of types of validity and data interpretation

- "...the emphasis here is on issues of meaning in measurement and of values in evaluation, but attention is also addressed to the role of values in measurement and of meaning in evaluation."

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

- Original source of "Nedelsky Method" for standard-setting

- Based on judges successively eliminating response choices for each item on a test that an F, D, C, B, A-student should be able to reject as wrong

- Paper contains detailed instructions on how to apply the procedure.

Poggio, J.P. (1984). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. [ERIC Document 249 267]

- Discusses educational competency tests (reading and math)

- Reviews experience with Angoff, Ebel, Nedelsky, Contrasting groups, and Borderline groups methods using panels of judges or survey-type questionnaires sent to large samples of judges

- Contrasting and Borderline Groups Method
    - contrasting group method easier
    - both tend to yield lower standards than Angoff or Ebel method

- Nedelsky Method
    - confusing for judges;
    - judges report not being confident in their judgments
    - can only be used by experienced teachers (read: experienced, relevant professionals)
    - yields lowest standard of all the methods

- Angoff Method
    - easy to implement and understand in panel or survey format
    - considerable variability among individual judge's standards (may be correctable if actual test response data is made available to the judges and/or if a Delphi procedure is used)
    - many judges have a problem defining a minimally competent student (could be corrected with agreed upon operational, i.e., behavioral, definition)

- Ebel Method
    - time-consuming (fatigue, boredom)
    - rather easy
    - problem with making ratings of "Questionable"...causes judges to become concerned about the method
    - computation of standard varies considerably depending on whether it is computed by judge or group cell values

- Kansas process now use Ebel and Angoff methods with survey format, this reference provides examples of the survey materials used
- Author cautions that while objective to a point, standard-setting is still value laden

Popham, W.J. (1978). As always, provocative. Journal of Educational Measurement, 15, 297-300.

- Critique of Glass (1978) position on standard-setting

- Suggests one way of looking at a definition of minimum competence is as "the lowest of proficiency which they [educators] consider acceptable for the situation at hand"..." 'lowest acceptable performance' conception of minimum competence."

Reid, J.B. (1984). Adapting a mandated pre-set pass/fail point. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document 246-069)

Abstract

While standard setting procedures are typically discussed in terms of deriving a reasonable cutting score for a given form of a test, the situation may be structured such that the standard has been mandated without regard to the test form itself. This situation may result either through legislative or policy actions and may be a fait accompli by the time someone experienced in standard setting methodology becomes involved. Although this may be an undesirable situation, it is not necessarily an impossible position from which to operate. This paper will explore an approach whereby an item "cut-score index" is included as an additional dimension in the test specifications for use in tailoring a test form to correspond to a pre-set pass/fail point. Issues such as the stability of such item index and the effect on content representativeness are discussed. (Author)

- Pass/fail point of 50% vs. 70% may not really reflect stringency of standards, but may simply reflect difference in difficulty levels of the items on the test...more a function of the distracters (response choices) than the knowledge tested itself

- Approach assumes existence of an item bank with relevant items possessing a range of item difficulties

- Can then construct a form of the test from items to obtain a particular mean cut-score index for the items, thereby tailoring a test to a given standard

- "To the extent that relevance influences judgments, the item's difficulty should not be expected to be perfectly related to the cut-score index, the difficulty index will provide only a very imperfect way of tailoring a test to a given standard. A clarification to judges on this point may help to reduce variability of judgments on individual items and increase the observed relationship between item difficulty and cut-score indices"

- Describes how the Nedelsky method would tend to depress overall standards and how the Angoff method tends to have the opposite effect; Ebel method requires judgments of item difficulty and item relevance, this is a potential short-coming of the method, since there is doubt regarding judges capability to make judgments on even a single dimension

- Suggests: 1) Using an item average cut-score from Nedelsky and Angoff methods to balance the bias; 2) panel of judges should represent a variety of interested parties to capture different perspectives (educators, entry-level job incumbents, supervisors, administrators, consumers, etc.) for each item in the item-bank; 3) number of judges, more=better; 4) verifying standards on an assembled test form: if a separate panel is to rate the test form, extra items should be included with a range of cut-score indices so that items may be selectively replaced to arrive at the desired standard (context of an item on assembled test may influence its perceived difficulty...and item of medium difficulty surrounded by hard items may seem easier than it really is).

Scriven, M. (1978). How to anchor standards. *Journal of Educational Measurement*, *15*, 273-276.

- Distressed that anyone in the educational test and measurement field would "talk as if R&D never existed, as if one doesn't *automatically* have to investigate the consequences of proposed standards in order to modify them for a second iteration, and so on... it suggests a lack of systems thinking, of self-evaluation, of knowledge about evaluation *as distinct from* testing that is extremely disturbing."

- Need: better procedures for calibrating and training judges, for synthesizing subtest scores, and in needs assessment...what are the needs to which a *test* is supposed to respond...the needs assessment should/would influence, e.g., the definition of "minimum acceptable standards"...also need to look at what the *skills* are needed for, e.g., would an 80% cut-score pass students who will later have trouble in a job?...need to "back track from problems later to 'needs now'...When you know how graduates need to perform on the job (the needs assessment) and you have a test you can use on pregraduates which has some predictive validity against job performance, you can set cutting scores (or bands)."

Shephard, L.A. (1984).. Setting performance standards. In R.A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 169-198). Baltimore, MD: John Hopkins University Press.

- Overview: describes basic methods and applications

- Competency-type standards are arbitrary in that they impose an artificial dichotomy...but that does not mean that they are capricious

- Regarding standard-setting procedures based on judgments of test content (Angoff, Ebel, Jaeger, Nedelsky)...prefers Angoff because it is the most straightforward; techniques differ only in how the rating task is posed to the judges (but this does, still, influence the standard)

- Use variety of judges...but there are pros and cons to having judges reach a consensus on a definition of minimal competence

- No compelling reason to use Borderline Groups method over Contrasting Groups procedure

- "Standards and expectations evolve from experience with typical rather than exceptional performance"

- Issue: how many objectives should be mastered to pass a course?; as an alternative, set standards to denote "mastery" but not make mastery of one subdomain prerequisite to the next

- Elements for a composite standard-setting model:
    1. obtain absolute judgments based on inspection of test questions
    2. perform an empirical validation with data based on judged masters and nonmasters
    3. obtain decisions about acceptable passing rates
    4. make adjustments for unreliability to minimize the costs of classification errors

-"If the test cannot discriminate in the region of the intended cut-off score, statistical adjustments (decision-theoretic error models) may lead to a bizarre policy (e.g., 100% passing). When this happens, the test should be revised rather than juggling the standard."

-"Apparent strengths and weaknesses (overall pass/fail rates) should never be interpreted from standards alone, without confirmation from normative comparisons."

Shephard, L.A. (1983).. Standards for placement and certification. In S.B. Anderson & J.S. Helmick (Eds.), On educational testing (pp. 61-90). San Francisco, CA: Jossey-Bass Publishers.

- Overview of issues and methods

- "...there is no way for the user to anticipate what philosophical or conceptual differences in the understanding of minimal competence will be reflected in the different operationalizations"

- Regarding supplemental use of normative data in standard-setting processes, "empirical data will make absolute deliberations better informed and more realistic...normative data are like validity evidence that can be used to cross-check goals set on the basis of test content"

- Three different uses of cut-scores: 1) Pupil placement decisions in the classroom 2) Certification of individuals 3) Program evaluation

- Different methods of standard-setting result in different standards...no single method is the most valid or logically correct

- Page 78: mention external standard and utility functions as a method for adjusting standards..."object is to match the test dichotomy to the criterion dichotomy to ensure that the smallest possible number of classification errors will be made"...Glass (1978) refers to these methods as "bootstrapping on other criterion scores" and criticizes the method because it accepts at face value the already existing standard (e.g., a required 60% correct for passing an SQT, to set an aptitude area cutoff score)...[this may be a particularly serious problem in education competency testing]...regarding utility functions in the form of loss ratios (cost of one error in relation to cost of another error [false positive vs. false negative]) practical usefulness is questioned ...assumes standard-setting problem has been solved for the criterion variable and that the judges will know how to assign the necessary utilities and choose the right shape for the utility function

- "How elaborate this standard-setting process should be will depend on how serious the consequences of invalid standards are."

Shephard, L.A. (1983).. Standards for placement and certification. In S.B. Anderson & J.S. Helmick (Eds.), On educational testing (pp. 61-90). San Francisco, CA: Jossey-Bass Publishers.

- Overview of issues and methods

- "...there is no way for the user to anticipate what philosophical or conceptual differences in the understanding of minimal competence will be reflected in the different operationalizations"

- Regarding supplemental use of normative data in standard-setting processes, "empirical data will make absolute deliberations better informed and more realistic...normative data are like validity evidence that can be used to cross-check goals set on the basis of test content"

- Three different uses of cut-scores: 1) Pupil placement decisions in the classroom 2) Certification of individuals 3) Program evaluation

- Different methods of standard-setting result in different standards...no single method is the most valid or logically correct

- Page 78: mention external standard and utility functions as a method for adjusting standards..."object is to match the test dichotomy to the criterion dichotomy to ensure that the smallest possible number of classification errors will be made"...Glass (1978) refers to these methods as "bootstrapping on other criterion scores" and criticizes the method because it accepts at face value the already existing standard (e.g., a required 60% correct for passing an SQT, to set an aptitude area cutoff score)...[this may be a particularly serious problem in education competency testing]...regarding utility functions in the form of loss ratios (cost of one error in relation to cost of another error [false positive vs. false negative]) practical usefulness is questioned ...assumes standard-setting problem has been solved for the criterion variable and that the judges will know how to assign the necessary utilities and choose the right shape for the utility function

- "How elaborate this standard-setting process should be will depend on how serious the consequences of invalid standards are."

Shephard, L.A. (1976). Setting standards and living with them. Paper presented to the National Council on Measurement in Education, San Francisco, CA.

- Defines "standard" as requiring an absolute, not relative, judgment of performance; therefore "criterion score" is the same as "performance standard"...the standard is the level of proficiency that each student is expected to attain; a separate issue is how to determine the level of performance that constitutes mastery

- Counter examples will always exist, but lowering standards until everyone passes defeats the purpose of standards

- Standard-setting is subjective, not capricious

- Harshness of standards ought to be modified depending upon the relative seriousness of false positives and false negatives

- Standard-setting ought to be an iterative process

- Normative (experiential) basis of judgments ought to be a formal part of the standard-setting process

- Suggests having more than one group of expert judges that meet separately...if similar standards are then reached the result will be more dependable (and defensible)

- "If standards are to be used to make decisions about individuals then one set of criteria is needed. The criterion should either be the most stringent or the most lenient of those proposed depending on which type of error is more serious in that situation. If false negatives would be more costly to individuals and society, then the standards should be lower than in the instances when false positives must be screened out."

Shikiar, R. & Saari, L.M. (1985). Establishing cut scores for the NRC reactor operator and senior reactor exam (Technical Evaluation Report No. PNL-5131). Seattle, WA: Pacific Northwest Laboratory.

- General review of issues and discussion of alternatives available to the Nuclear Regulatory Commission (NRC)

- NRC operator exams are unusual in that they are partially oral exams and the non-written tests are scored pass/fail

- Given policy and public confidence concerns, a high cut-off score (80%) is preferred, permitting (requiring) adjustment of test content

- [Interesting report from the perspective of standard-setting issues in public sector, high visibility (in terms of being open to public scrutiny and concern) jobs, as opposed to educational competency testing.]

TOWARD A GENERAL MODEL OF SOLDIER EFFECTIVENESS: FOCUSING ON THE COMMON ELEMENTS OF PERFORMANCE

Walter C. Borman
Elaine D. Pulakos
Stephen J. Motowidlo

Personnel Decisions Research Institute

Presented on symposium,
"Multiple Criteria and Multiple Jobs: Will One Model Fit All?"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Toward a General Model of Soldier Effectiveness:
Focusing on the Common Elements of Performance

The Army Research Institute for the Behavioral and Social Sciences (ARI) initiated Project A, a nine-year research program intended to link selection and classification standards to job performance. The primary goal of Project A is to achieve increased Army effectiveness through improving the soldier-job match. This goal will be accomplished by developing a comprehensive set of selection and classification measures (predictors) and performance criteria, and empirically investigating relationships between these predictor and performance measures.

As part of the effort, development proceeded on "Army-wide" rating dimensions, elements of soldier effectiveness that might be relevant for first-term soldiers in *any* U.S. Army military occupational specialty (MOS). A previous report related to this objective of developing Army-wide dimensions described a conceptual model of soldier effectiveness and a scale development effort to establish empirically derived dimensions of Army-wide effectiveness (Borman, Motowidlo, & Hanser, 1983). The present paper (1) reviews briefly the earlier conceptual model of soldier effectiveness and the Army-wide scale development results; (2) describes a large scale administration of the scales to peer and supervisor raters of over 8000 soldiers in 19 MOS; and (3) details factor analysis findings for these performance ratings and discusses similarities in underlying structures of the ratings across the 19 MOS.

Developing a Conceptual Model of Soldier Effectiveness

Early in Project A we developed a conceptual or theoretical model of soldier effectiveness. This was an effort to lay out a hypothetical framework depicting the performance requirements for first-term soldiers, what it might take to be an all-around effective performer during first-term in the Army.

In this model-building effort, we sought to define a set of performance-related factors that would include elements of soldier effectiveness not directly related to task performance, but related instead to a broader conception of job performance. We believed that being a good soldier from the Army's perspective means more than just performing the job in a technically proficient manner. It also means performing a variety of other activities that contribute to a soldier's effectiveness in the unit and to his or her "overall worth to the Army." Our preliminary model presumed that soldier effectiveness could be analyzed according to the elements that comprise the constructs of organizational commitment, organizational socialization, and morale.

Briefly, the first construct, organizational commitment, refers to tne strength of a person's identification with and involvement in the organization. It incorporates three kinds of elements: acceptance and internalization of organizational values and goals; motivation to exert effort toward the accomplishment of organizational objectives; and firm intentions of staying in the organization. Organizational commitment involves a sense of loyalty to the organization as a whole and a desire to fulfill more general role requirements that come with organizational membership.

The second construct, organizational socialization, refers to the process an organization member goes through to acquire the social knowledge and skills necessary to assume a useful organizational role. When the socialization process is successful, a person will acquire not only job-related skills, but also new patterns of behavior with subordinates, peers, and superiors in the organization, new attitudes, beliefs, and values in line with organizational norms. Such individual changes are frequently crucial for assuring that the behaviors of different individual organization members will be smoothly coordinated toward accomplishing the organization's mission.

The concept of morale (the third construct in the model) has traditionally been seen as extremely important in military organizations. Morale is multifaceted. It involves feelings of determination to overcome obstacles, confidence about the likelihood of success, exaltation of ideals, optimism even in the face of severe adversity, courage, discipline, and group cohesiveness. In one study designed to identify behavioral dimensions of morale in the U.S. Army, the following dimensions were found to efficiently describe behavioral expressions of morale among soldiers: community relations; teamwork and cooperation; reactions to adversity; superior-subordinate relations; performance and effort on the job; bearing, appearance, marching, and military courtesy; pride in unit, Army, and country; and self-development during off-duty hours. Because morale seems to figure so prominently as a determinant of unit effectiveness, behavioral dimensions like these may also in part represent important elements of individual soldier effectiveness.

These three broad constructs can be viewed in another way that leads to a more concrete view of soldier effectiveness. From the combination of morale and commitment emerges a general category that can be labeled "Determination." It is a motivational category that reflects the spirit, strength of character, or "will-do" aspects of good soldiering. Morale and socialization lead to "Teamwork," behaviors that have to do with effective relationships with peers and the unit. Commitment and socialization give rise to "Allegiance." This taps into acceptance of Army norms with respect to authority, faithful adherence to orders, regulations, and the Army lifestyle, and being adjusted and socialized to the point of wanting to continue in the soldiering role and stay in the Army.

Each general category of effectiveness subsumes five more specific dimensions. These dimensions were developed and defined according to our preliminary expectations of how the elements implied by determination, teamwork, and allegiance might suggest specific behavioral patterns of soldier effectiveness. The preliminary conceptual model is summarized in Figure 1.

The Empirical Model

The conceptual model provides interesting hypotheses about first-term soldier performance requirements. However, we believed strongly that follow-up work was required to identify more concretely the performance factor domain. An excellent strategy to accomplish this is the behaviorally-anchored rating scale or BARS method (Smith & Kendall, 1963; Campbell, Dunnette, Arvey, & Hellervik, 1973). Using this method, persons knowledgeable about the target job write vignettes or stories relating actual first-term soldier job-related behavior in which incumbents on the job performed effectively, in the middle range, or ineffectively. In the present research, 30 NCOs (mostly E-5 to E-7), and 47 officers (mostly Captains and Majors) in many different MOSs and specialty areas stationed in four different locations generated a total of

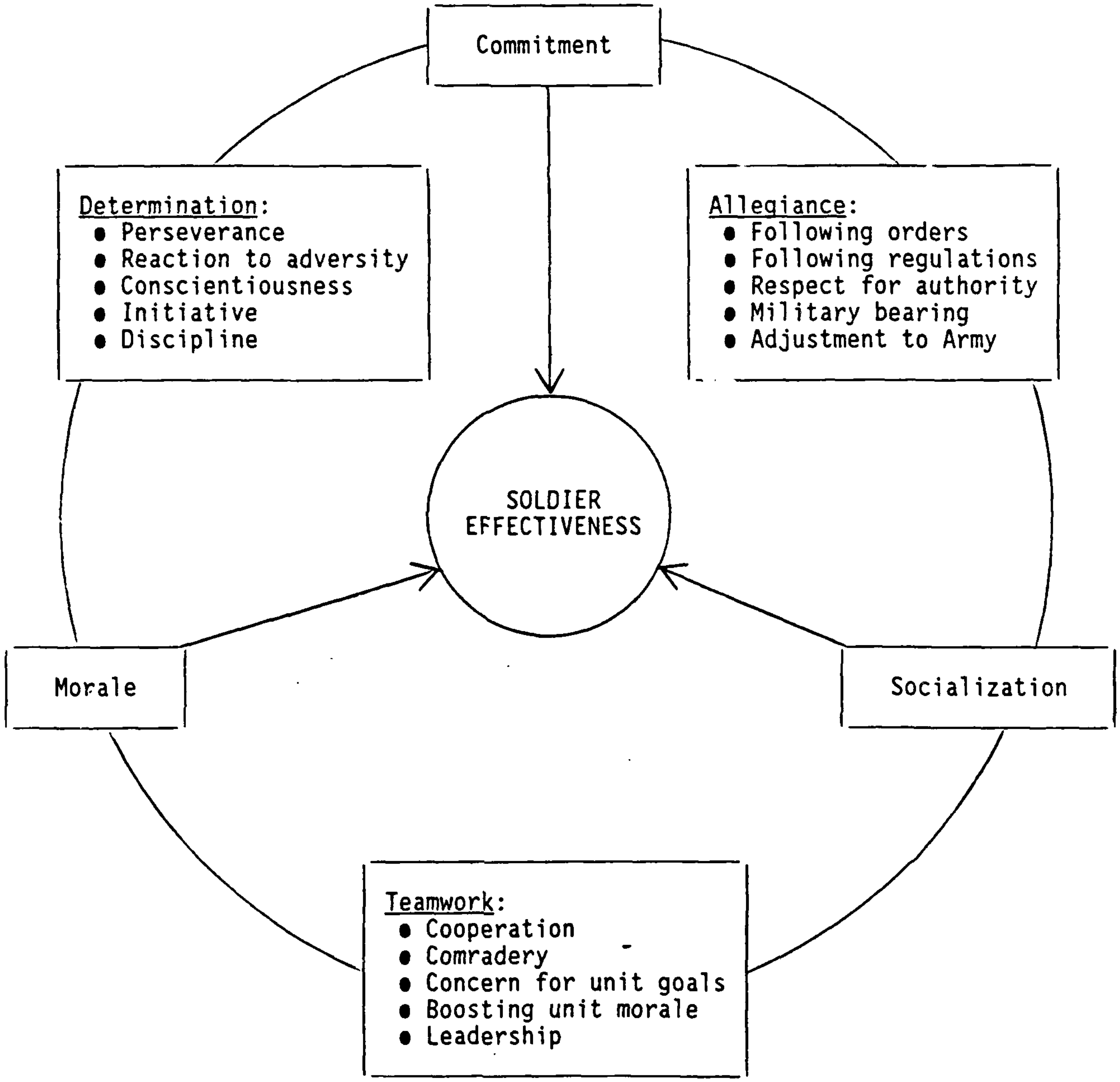


Figure 1. A Preliminary Model of Soldier Effectiveness

1315 performance examples reflecting all the various elements of soldier performance. Two performance examples are provided here to give an idea of what they look like.

- Although this soldier knew there was an NBC room inspection coming up, he did not put the masks in order or empty the trash cans in the room (relatively ineffective).
- When ordered to inventory three magazines, this soldier accomplished the task without supervision during his lunch time (relatively effective).

The 1315 performance examples were examined closely by our research staff, and 13 categories or dimensions of performance were formed based on the content of these many examples. Figure 2 contains examples of these categories.

Next, the performance examples were retranslated. Sixty NCOs and officers sorted each example into one of the categories according to its content and rated the effectiveness level it reflected. Seventy-eight percent of the performance examples were sorted into a single dimension by more than 50% of retranslation raters <u>and</u> had standard deviations of less than 2.0 on a 9-point scale. These findings indicate that many of the examples were unambiguous relative to the performance category into which they were sorted and the effectiveness level represented. The category system was revised based on retranslation results, with two pairs of categories being combined and one category subsequently dropped from further consideration. Thus, 10 effectiveness categories, appearing in Figure 2, represent the empirical model of soldier effectiveness.

A. *Technical Knowledge/Skill - Displaying job and soldiering knowledge/skill.*

B. Effort - Showing initiative and extra effort on the job/mission/assignment.

C. Following Regulations and Orders - Adhering to regulations, orders, and SOP and displaying respect for authority.

D. Integrity - Displaying honesty and integrity in job-related and personal matters.

E. Leadership - Performing in a leader role, as required, and providing guidance and support for fellow unit members.

F. Maintaining Assigned Equipment - Checking on and maintaining own weapon/vehicle/other equipment.

G. Military Appearance - Maintaining proper military appearance.

H. Physical Fitness - Maintaining military standards of physical fitness.

I. Self-Development - Developing own job and soldiering skills.

J. Self-Control - Controlling own behavior related to drugs/alcohol and aggressive acts.

Figure 2. Empirical Category System for the Model of Soldier Effectiveness

Behavior based rating scales were developed to measure performance on these 10 dimensions. In addition, a rater orientation and training program was prepared (Pulakos & Borman, 1985) to guide peer and supervisor raters in completing the ra:ings accurately. The resulting Army-wide rating system was then field-tested with peers and supervisors of first-term soldiers in nine different MOSs. Approximately 3 peers and 2 supervisors of each of 150 soldiers per MOS used the rating package to evaluate the effectiveness of their peers or subordinates. Results of these field tests relevant for the present paper suggest that the dimensions are in fact appropriate for measuring soldier effectiveness across different MOSs. Raters in all nine MOSs were able to complete the scales. The distributions of peer and supervisor ratings were reasonable (almost always a mean between 4.5 and 4.75 on a 7-point scale and standard deviations of between .90 and 1.15), and the intraclass correlation interrater reliabilities on individual scales were around .60 for peers and for supervisors.

The rating scales were revised slightly based on field test feedback and readied for administration to peer and supervisor raters of first-term soldiers in 19 MOSs during the concurrent validation (CV) data collection. The purpose of this paper is to characterize the dimensional structure of the resulting peer and supervisor ratings and to compare structures across the 19 MOSs. This comparison will help to assess further similarities (and differences) in the ways the Army-wide rating scales are used in different MOSs.

METHOD

Table 1 presents the numbers of raters and ratees in the CV sample. "Batch A" refers to MOSs which received a larger number of performance measures (including job sample and knowledge tests) than did the "Batch

Table 1

Concurrent Validation Sample

	Peers			Supervisors		
MOS	Number of Ratees (Soldiers)	Total Number of Ratings	Rater/ Ratee Ratio	Number of Ratees	Total Number of Ratings	Rater/ Ratee Ratio
Batch A						
11B	679	2,377	3.50	650	1,242	1.92
13B	633	2,204	3.48	638	1,218	1.91
19E	485	1,601	3.30	490	934	1.91
31C	316	856	2.71	349	637	1.83
63B	559	1,467	2.62	597	1,158	1.94
64C	646	2,396	3.71	639	1,206	1.89
71L	422	990	2.35	460	788	1.71
91A	481	1,551	3.23	468	954	2.04
95B	681	2,543	3.73	652	1,255	1.92
Total (A)	4,902	15,985	3.26	4,943	9,392	1.90
Batch Z						
12B	684	2,325	3.40	672	1,248	1.86
16S	461	1,670	3.62	377	782	2.07
27E	141	454	3.22	143	271	1.90
51B	100	263	2.63	104	196	1.88
54E	372	1,139	3.06	372	649	1.74
55B	271	829	3.06	264	437	1.66
67N	265	867	3.27	245	421	1.72
76W	422	1,215	2.88	419	803	1.92
76Y	454	836	1.85	548	916	1.67
94B	570	1,168	2.94	546	1,030	1.89
Total (Z)	3,740	10,766	2.88	3,690	6,753	1.83
Total (A and Z)	8,642	26,751	3.10	8,633	16,145	1.87

NOTE:

11B - Infantryman
13B - Cannon Crewman
19E - Tank Crewmember
31C - Radio Operator
63B - Vehicle Mechanic
64C - Motor Transport Operator
71L - Administrative Specialist
91A - Medical Specialist

95B - Military Police
12B - Combat Engineer
16S - MANPADS Crewman
27E - Chemical Operations Specialist
51B - Carpentry Masonry Specialist
54E - Chemical Operations Specialist

55B - Ammunitions Specialist
67N - Utility Helicopter Repair
76N - Petroleum Supply Specialist
76Y - Unit Supply Specialist
94B - Food Service Specialist

Z" MOSs. As can be seen, more than 8000 first-tour soldiers were rated by about 3 peers and 2 supervisors (per ratee) in the 19 MOSs represented. As in the field tests, individual peer rating sessions were held with approximately 15 first-tour soldiers, and supervisor sessions were conducted with generally 5-15 supervisors of the target ratees. In each session, raters went through the rater orientation and training program and then completed their evaluations.

Data analyses were first concerned with interrater agreement. Intraclass correlations were computed to index the reliability of ratings within and across rating source (i.e., peer and supervisors). Second, the structure of peer and supervisor ratings within each MOS was examined. Ratings within rating source, but across MOSs on individual dimensions were intercorrelated and factor analyzed. Then, MOS by MOS, dimension ratings were correlated with factor scores representing each of the factors to investigate consistency across MOSs in the patterns of dimension-factor relationships.

RESULTS

Table 2 contains interrater reliability information for these ratings within rating source and MOS on the Army-wide scales. Reliabilities are reasonably high on individual dimensions, right around .50 on average for both peers and supervisors. A unit weighted composite of the rating scales provides higher reliabilities, as noted in the tables. Intraclass correlations depicting reliabilities of the unit weighted composite across the peer and supervisor sources range from .36 to .54 and average .48, lower than within source, but still reasonably high. Incidentally, reliabilities of ratings on these Army-wide scales are consistently higher than comparable reliabilities of

TABLE 2

INTERRATER RELIABILITIES

BATCH A MOS, PEERS

	11B	13B	19E	31C	63B	64C	71L	91A	95B	Mean
Scale Range	41 - 61	35 - 60	38 - 63	30 - 61	27 - 54	40 - 60	18 - 42	36 - 66	35 - 71	
Median Scale	54	49	47	44	48	49	34	51	52	
Unit-Weighted Composite	63	61	62	57	54	61	41	59	63	58

BATCH A MOS, SUPERVISORS

	11B	13B	19E	31C	63B	64C	71L	91A	95B	Mean
Scale Range	48 - 61	36 - 59	45 - 59	40 - 64	48 - 69	41 - 62	40 - 68	51 - 62	33 - 61	
Median Scale	58	48	51	49	57	51	54	56	49	
Unit-Weighted Composite	69	54	64	64	65	65	72	68	60	65

BATCH Z MOS, PEERS

	12B	16S	27E	51B	54E	55B	67N	76W	76Y	94B	Mean
Scale Range	43 - 61	48 - 66	47 - 68	35 - 69	39 - 67	15 - 48	49 - 68	28 - 50	05 - 44	24 - 50	
Median Scale	54	58	51	55	53	39	55	39	22	45	
Unit-Weighted Composite	65	69	71	67	64	45	68	46	33	53	58

BATCH Z MOS, SUPERVISORS

	12B	16S	27E	51B	54E	55B	67N	76W	76Y	94B	Mean
Scale Range	40 - 61	46 - 66	36 - 65	46 - 75	42 - 66	42 - 65	36 - 61	35 - 55	38 - 57	42 - 53	
Median Scale	49	57	55	61	53	52	48	45	45	50	
Unit-Weighted Composite	61	68	70	81	66	66	59	56	56	60	64

ratings on MOS-specific behavior-based job performance scales (.41 and .49, respectively for peer and supervisor unit weighted composites). For the total sample, ratings on the 10 dimensions were intercorrelated and then factor analyzed using the principal factor method with communality estimates in the diagonal and varimax rotation. A three-factor solution emerged for both sources. The factors were named: (1) Job Skills and Motivation, (2) Discipline, and (3) Personal Appearance. Solutions for the peer and supervisor rating sources appear in Table 3. Next, for each rating source and MOS separately, ratings on individual dimensions were correlated with factor scores for each of the three factors (as mentioned in the Method Section). Thus, Tables 4 and 5 summarize the similarity between different rating sources and MOSs in the factor structure of the Army-wide ratings. The same analysis was conducted for the pooled peer and supervisor ratings, and those results appear in Table 6.

Findings in the three tables for the most part support the stability and appropriateness of the three-factor structure across rating source and MOS. We checked correlations between dimension ratings and factor scores for each rating source-by-MOS combination to identify instances where dimension ratings related higher with a factor other than the one they were supposed to correlate highest with according to the across-MOS results. For peer ratings, Maintaining Equipment shifts back and forth between Factors 1 and 3. For 7 of the 19 MOSs, correlations between that dimension's ratings and Factor 3 are higher than the correlations with Factor 1. Conceptually, this is not too troublesome because Maintaining Equipment might be seen as, rather than a core technical skill or motivation-related dimension (Factor 1), a more peripheral, appearance or maintenance-oriented dimension (Factor

Table 3

Factor Analysis Results for Peer and

Supervisor Ratings on Army-Wide Scales

Peer Ratings			
	Rotated Factor Patterns		
Dimensions	Factor 1	Factor 2	Factor 3
Technical Skill	.65	.30	.32
Leadership	.62	.32	.40
Effort	.60	.45	.29
Self Development	.49	.40	.38
Maintain Equipment	.43	.35	.39
Following Regulations	.34	.65	.28
Self-Control	.20	.57	.19
Integrity	.43	.55	.29
Military Appearance	.28	.31	.56
Physical Fitness	.22	.15	.50

NOTE: Percent of common variance is 39, 35, and 26, respectively, for the three factors.

Supervisor Ratings			
	Rotated Factor Patterns		
Dimensions	Factor 1	Factor 2	Factor 3
Technical Skill	.70	.26	.30
Leadership	.68	.30	.34
Effort	.70	.40	.25
Self Development	.55	.34	.39
Maintain Equipment	.53	.32	.38
Following Regulations	.42	.66	.30
Self-Control	.22	.61	.23
Integrity	.49	.57	.30
Military Appearance	.33	.32	.55
Physical Fitness	.20	.17	.47

NOTE: Percent of common variance is 46, 31, and 23, respectively, for the three factors.

Table 4

CORRELATIONS BETWEEN ARMY-WIDE BEHAVIORAL FACTORS AND ARMY-WIDE BEHAVIORAL DIMENSIONS FOR PEER RATERS ACROSS MOS

	Factor 1	Factor 2	Factor 3
Factor 1			
Factor 2	.39 (.23-.51)		
Factor 3	.48 (.36-.59)	.34 (.16-.51)	
Technical Knowledge	.87 (.83-.91)	.40 (.27-.50)	.50 (.36-.61)
Leadership	.83 (.79-.88)	.43 (.31-.52)	.61 (.53-.70)
Effort	.81 (.74-.86)	.61 (.50-.71)	.44 (.35-.56)
Self-Development	.66 (.54-.77)	.55 (.35-.65)	.60 (.32-.75)
Maintaining Equipment	.59 (.42-.69)	.48 (.37-.60)	.61 (.53-.68)
Following Regulations	.46 (.35-.60)	.89 (.85-.91)	.43 (.28-.56)
Self-Control	.27 (.04-.36)	.78 (.72-.84)	.29 (.20-.39)
Integrity	.58 (.47-.67)	.75 (.72-.81)	.45 (.27-.59)
Military Appearance	.38 (.18-.51)	.42 (.30-.56)	.87 (.83-.90)
Physical Fitness	.30 (.12-.40)	.21 (-.02-.42)	.77 (.71-.81)

NOTE: In parentheses is the range of correlations that resulted for individual MOS. The numbers not in parentheses are total sample correlations between the dimension ratings and factor scores.

Factor 1: Job Relevant Skills and Motivation
Factor 2: Personal Discipline
Factor 3: Personal Appearance

Table 5

CORRELATIONS BETWEEN ARMY-WIDE BEHAVIORAL FACTORS AND ARMY-WIDE BEHAVIORAL DIMENSIONS FOR SUPERVISOR RATERS ACROSS MOS

	Factor 1	Factor 2	Factor 3
Factor 1			
Factor 2	.30 (.13-.36)		
Factor 3	.39 (.27-.51)	.34 (.23-.47)	
Technical Knowledge	.86 (.83-.89)	.35 (.19-.42)	.46 (.34-.54)
Leadership	.84 (.81-.92)	.40 (.22-.49)	.53 (.44-.62)
Effort	.87 (.83-.89)	.53 (.40-.59)	.39 (.25-.48)
Self-Development	.67 (.61-.73)	.45 (.27-.50)	.62 (.55-.68)
Maintaining Equipment	.66 (.54-.71)	.42 (.30-.52)	.60 (.51-.72)
Following Regulations	.52 (.41-.58)	.88 (.86-.90)	.48 (.38-.56)
Self-Control	.27 (.13-.36)	.81 (.78-.84)	.36 (.27-.44)
Integrity	.60 (.48-.65)	.76 (.63-.80)	.47 (.34-.57)
Military Appearance	.40 (.31-.50)	.43 (.28-.52)	.87 (.84-.90)
Physical Fitness	.24 (.15-.36)	.22 (.12-.37)	.74 (.68-.81)

Note: In parentheses is the range of correlations that resulted for individual MOS.

Factor 1: Job Relevant Skills and Motivation
Factor 2: Personal Discipline
Factor 3: Personal Appearance

Table 6

CORRELATIONS BETWEEN ARMY-WIDE BEHAVIORAL FACTORS AND ARMY-WIDE BEHAVIORAL DIMENSIONS FOR COMBINED PEER AND SUPERVISOR RATERS ACROSS MOS

	Factor 1	Factor 2	Factor 3
Factor 1			
Factor 2	.28 (.13-.37)		
Factor 3	.40 (.30-.51)	.28 (.13-.36)	
Technical Knowledge	.88 (.85-.91)	.36 (.25-.45)	.46 (.32-.58)
Leadership	.85 (.82-.89)	.38 (.27-.47)	.58 (.50-.64)
Effort	.85 (.80-.88)	.55 (.44-.65)	.41 (.31-.49)
Self-Development	.70 (.63-.79)	.48 (.36-.55)	.59 (.47-.75)
Maintaining Equipment	.67 (.59-.74)	.43 (.36-.54)	.55 (.47-.67)
Following Regulations	.50 (.41-.63)	.89 (.86-.91)	.46 (.36-.56)
Self-Control	.27 (.08-.35)	.81 (.79-.85)	.31 (.15-.36)
Integrity	.61 (.46-.70)	.76 (.71-.81)	.44 (.34-.53)
Military Appearance	.39 (.23-.49)	.41 (.32-.53)	.89 (.75-.92)
Physical Fitness	.25 (.14-.37)	.18 (-.03-.31)	.75 (.69-.80)

Note: In parentheses is the range of correlations that resulted for individual MOS.

Factor 1: Job Relevant Skills and Motivation
Factor 2: Personal Discipline
Factor 3: Personal Appearance

3). In addition, for 2 of the 19 MOS, ratings on Self Development correlate higher with Factor 3 than with Factor 1.

For the supervisor raters, four MOSs have ratings on Maintaining Equipment correlating higher on Factor 3 than on Factor 1, and one MOS has Self Development correlating as high with Factor 3 as with Factor 1. When ratings from the two sources are pooled (Table 5), ratings for only two MOSs on Maintaining Equipment correlate higher with Factor 3 than with Factor 1.

There are no other such "reversals" for individual rating sources and MOSs. For the vast majority of dimension-by-MOS combinations, the dimension correlates highest with the factor it "belongs to" according to factor analysis results for the entire sample. This is admittedly only one possible way to explore stability of the three-factor solution across MOSs and rating source, but the results do suggest that the solution represents a consistent interpretable three-factor summary of the 10 Army-wide dimensions.

DISCUSSION

The three Army-wide factors of (1) Job Skills and Motivation, (2) Discipline, and (3) Personal Appearance hold up reasonably well across rating sources and MOS.

For the vast majority of dimension-by-MOS combinations, correlations between dimension ratings and factor scores on the three summary factors reflect the same pattern across MOSs (<u>and</u> across the peer and supervisor rating sources). Of course, other methods of assessing similarity in performance requirements across MOSs may well yield different results. Behavior dimensions emerging from the BARS research are at a rather general level and emphasize the common performance requirements across MOSs. Specific task rating scales, for example, developed for each MOS would not reflect nearly so much the common ground across the different MOSs (yet might be quite important to include to round out a complete picture of the MOS's performance requirements).

Nonetheless, the conceptual model of soldier effectiveness and the empirical dimensions represent important effectiveness constructs that seem to be important for performing in any of the MOSs. Further, factor analytic work suggested that the structure of effectiveness on the 10 Army-wide dimensions might be quite similar across MOSs. That is, in most cases the ratings can be summarized by the same three-factor system across MOSs (and rating sources).

An important implication of these results is that the three summary factors presented here might serve as a common soldier effectiveness construct framework for all MOSs. Regarding criterion measurement in personnel selection research, there are advantages in having three constructs at a level of generality/specificity such that they reflect three <u>different</u> effectiveness criteria, but also are important in all MOSs. It should be possible with criterion measures tapping constructs at this level of generality to perform anaTyses of predictor-criterion relationships across MOSs (in addition to within MOS). Regarding the common framework notion, other criterion measures (besides the Army-wide performance ratings) might also be used to tap constructs within this framework.

REFERENCES

Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983). A model of individual performance effectiveness: Thoughts about expanding the criterion space. Paper in symposium, Integrated criterion measurement for large-scale computerized selection and classification, 91st Annual American Psychological Association Convention, Anaheim, CA.

Pulakos, E. D., & Borman, W. C. (1985). Rater orientation and training. In E. D. Pulakos & W. C. Borman (Eds.), Development and field test of Army-wide rating scales and the rater orientation and training program (Chpt. 5). Alexandria, VA: U.S. Army Research Institute Working Paper.

PROJECT A:
WHEN THE TEXTBOOK GOES OPERATIONAL

John P. Campbell

Human Resources Research Organization

Invited address presented at the Annual Convention of the American Psychological Association

Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

When the Textbook Goes Operational[1]

John P. Campbell
University of Minnesota

Titles for APA papers are usually chosen without careful thought and well in advance of their actually being written, often to the embarrassment of the speaker. This occasion is no exception. What I would like to talk about is the U.S. Army's Selection and Classification Project, or Project A for short, and some of its salient features and initial findings that I hope will be of interest. Also, even though I speak about it today, I am just one of over 20 investigators drawn from the Army Research Institute and three different research organizations. However, they should not be held responsible for these remarks.

We have often asked graduate students on preliminary exams to assume that they have been provided lots of resources and lots of time to do a really substantial research project and then to design such a project. That is, the instruction is to describe a fantasy relevant to their chosen area of interest. Students usually have difficulty outlining a project that is thorough enough and inclusive enough to use up all the hypothetical resources. The Army Selection and Classification Project (Project A) is a 9 year personnel research project with a total budget of over 20 million dollars. In many respects it is the prelim question's fantasy come true, and what happened on the way to reality is what I want to talk about for the next few minutes. After recounting just a bit of the project's history, I would like to: a) outline the objectives, b) briefly describe the project's organization and design, and c) summarize its major substantive activities and findings during its first 3 and one half years. Most importantly, I will try to highlight some of the special issues in research design, measurement, and prediction that perhaps only such a large project can address directly.

First let me say that we have had no problem spending the money or in filling the time. However, where there are benefits there are costs and along with being one of the largest personnel research projects ever undertaken, Project A is also one of the most closely monitored and reviewed. There are approximately 16 regularly scheduled review meetings each year. That is never part of the fantasy, nor is it ever mentioned in any textbook. We were unprepared for the amount of effort required for meetings and progress reports.

Now a bit of history. In the beginning there was a strong congressional mandate to validate the current DoD selection and classification tests against criteria of job performance. Most previous validation efforts had to rely on training achievement measures as criteria. There was a parallel desire on the part of the Army to examine whether additional selection tests should be developed to supplement the Armed Services Vocational Aptitude Battery (ASVAB). These two objectives fell upon the head and shoulders of an imaginative group of in-house professionals at the Army Research Institute (ARI) who decided to think big. Instead of a series of projects designed to address the two objectives why not examine the entire enlisted personnel selection and classifica-

[1] This research was funded by the Army Research Institute Contract No. MDA 903-82-C-0531. All statements expressed in this paper are those of the author and do not necessarily express the official positions of the U.S. Army Research Institute or the Department of the Army.

tion decision system with one integrated long term project. The idea was contrary to our usual way of doing research in industrial and organizational psychology where the typical model is a single investigator (or at the most two) who operates independently with a relatively small budget and a carefully delimited set of objectives. For reasons I do not really understand, the idea of a single project spanning several years with multiple objectives directed at an entire organizational system survived. In fact, the RFP specified a project that was simply too big for any one research firm or university group to undertake. As a result, coalitions formed to submit proposals. The contract was awarded, and the project begun on October 1, 1982 (FY83).

Objectives

Project A is directed at multiple operational and research objectives. The major ones are shown in Table 1.

The current selection classification system for enlisted personnel screens 300 to 400 thousand people each year, selects 120-140 thousand of them, and assigns each individual to one of approximately 275 entry level positions. The primary selection instrument is the Armed Services Vocational Aptitude Battery (ASVAB) which currently has ten subtests and four composites of subtests developed for different categories of occupational specialties. Cutting scores have been established for each job, or Military Occupational Speciality (MOS), and if the individual is above the cutting score on the appropriate ASVAB composite, assignments are made on the basis of Army needs, training space availability, and individual preferences. A system of bonuses is currently in use to influence individual preferences in the direction of Army needs.

The mandate of Project A is to develop an experimental battery of new selection/classification instruments, validate them against appropriate measures of job performance, assess their collective differential validity for doing "true" classification, and provide the information necessary for conducting "what if" games with differential weights for job assignments, changes in cutting scores, quotas, etc..

In the course of trying to meet this mandate, the Project has taken a very broad approach and has gone after a number of more basic scientific questions as well.

For example, we have tried to provide a systematic description, in a taxonomic sense, of the universe of information that is potentially useful for making predictions of future job performance and to develop a model of its latent structure. Similarly the Project has tried to develop a general latent structure model of job performance for entry level skilled jobs, at least as they are represented by the population of jobs performed by enlisted personnel in the U.S. Army.

If the latent structure of job performance and the taxonomic structure of selection/classification prediction information is modeled, and measures

Table 1

ARMY SELECTION AND CLASSIFICATION PROJECT

Operational Objectives

1) Develop new measures of job performance that can be used as criteria against which to validate selection/classification measures.

2) Validate existing selection measures against both existing and project-developed criteria.

3) Develop and validate new selection and classification measures.

4) Develop a utility scale for different performance levels across MOS.

5) Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility.

Research Objectives

1) Identify the constructs that constitute the universe of information available for selection/classification into entry level skilled jobs.

2) Develop a general model of performance for entry level skilled jobs.

3) Investigate the construct validity of the "method" variance in job performance measures.

4) Describe the utility functions and the utility metrics that individuals actually use when estimating "utility of performance."

5) Estimate the degree of differential prediction across (a) major domains of predictor information (e.g., abilities, personality, interests), (b) major factors of job performance, and (c) different types of jobs.

6) Determine the extent of differential prediction across racial and gender groups for a systematic sample of individual differences, performance factors, and jobs.

7) Develop new statistical estimators of classification efficiency.

are developed to assess the major constructs for a representative sample of jobs from a large population of jobs then a large set of interesting questions opens up. For example, to what extent is differential prediction possible across major components of performance? To what extent is there differential prediction across the major components of the predictor universe? To what extent does validity generalization across jobs depend upon the performance component being assessed? For any differential prediction across race and gender groups, what is the precise source of such differential regressions in terms of predictor component/performance component combinations? What happens to the overall regression line when those components are omitted?

Project Organization

The consortium that is working on Project A consists of psychologists drawn from three research organizations (HumRRO, AIR, and PDRI), the Army Research Institute, and a few stragglers such as myself. The Project's activities are divided into 4 substantive tasks corresponding to predictor development, training performance measurement, measurement of general job performance, and measurement of job specific performance components. There is also a major section devoted to data analysis and management of the centralized data base. Finally there is a management group that has the overall responsibility for the budget, schedules, coordination of substantive tasks, planning, reporting, interaction with advisory groups, etc..

There are three formally constituted advisory groups consisting of a general officers advisory group, a scientific advisory group, and a group made up of professional representatives from each of the services. These organizational arrangements are portrayed in Figures 1 and 2.

Basic Design

The basic design of Project A shown in Figure 3 is simply that of a very large test validation study that incorporates several independent data collections.

There are four major data files. The first consists of the available computer records for people who joined the Army in 1981 and 1982. The basic data are the ASVAB, the available training school grades, and the Skills Qualification Test (SQT), which is a paper-and-pencil measure of current job knowledge constructed, administered, and scored by the individual's unit command. Complete data were available on at least 100 people for only 98 of the 275 MOS, which simply illustrates another law of reality usually not found in textbooks. Using computerized files for any purpose other than the one for which they were designed will endanger the mental well being of most investigators.

FIGURE 1

Project A Organization

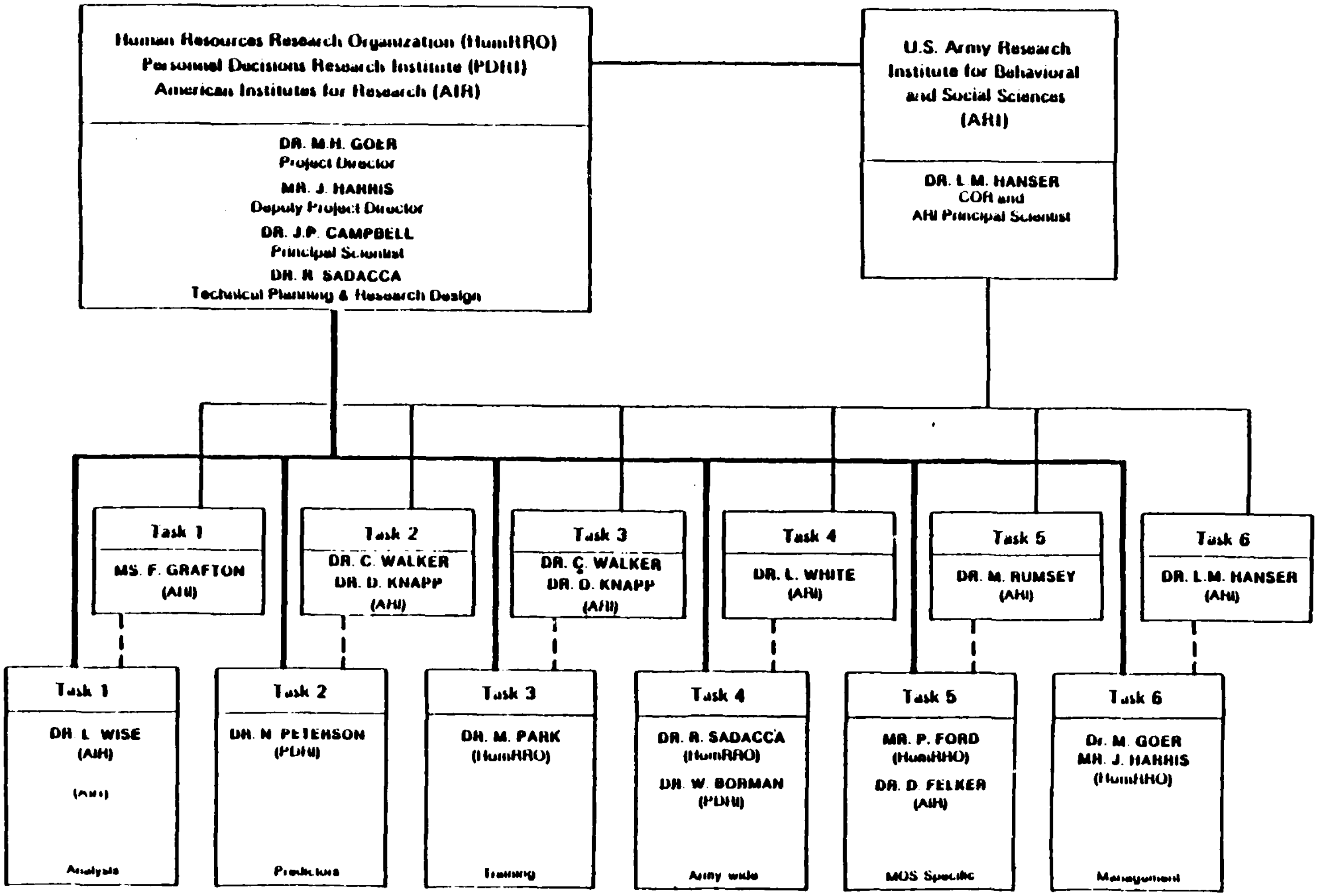
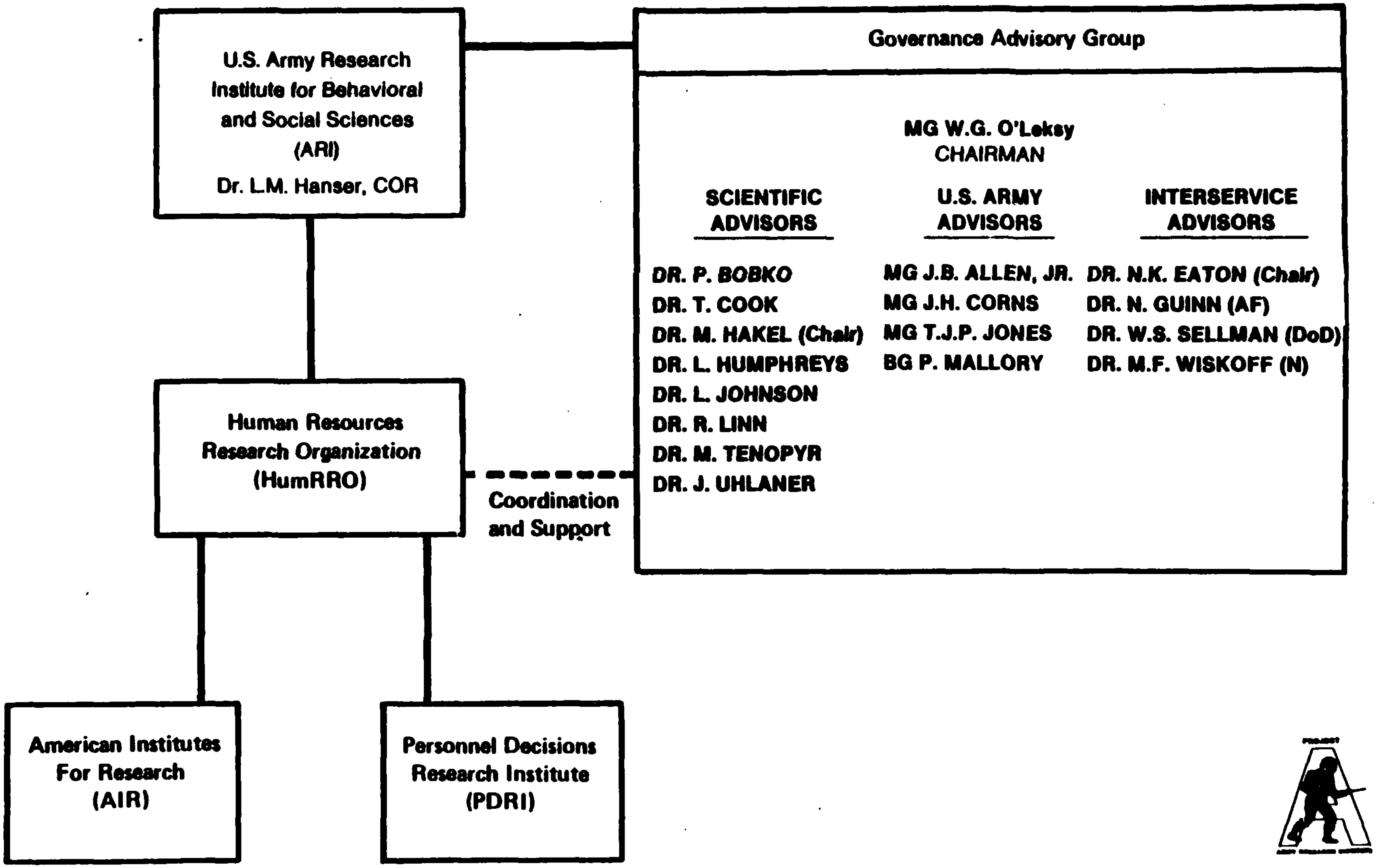


FIGURE 2

PROJECT A MANAGEMENT GROUP



U.S. Army Research Institute for Behavioral and Social Sciences (ARI)
Dr. L.M. Hanser, COR

Human Resources Research Organization (HumRRO)

Coordination and Support

American Institutes For Research (AIR)

Personnel Decisions Research Institute (PDRI)

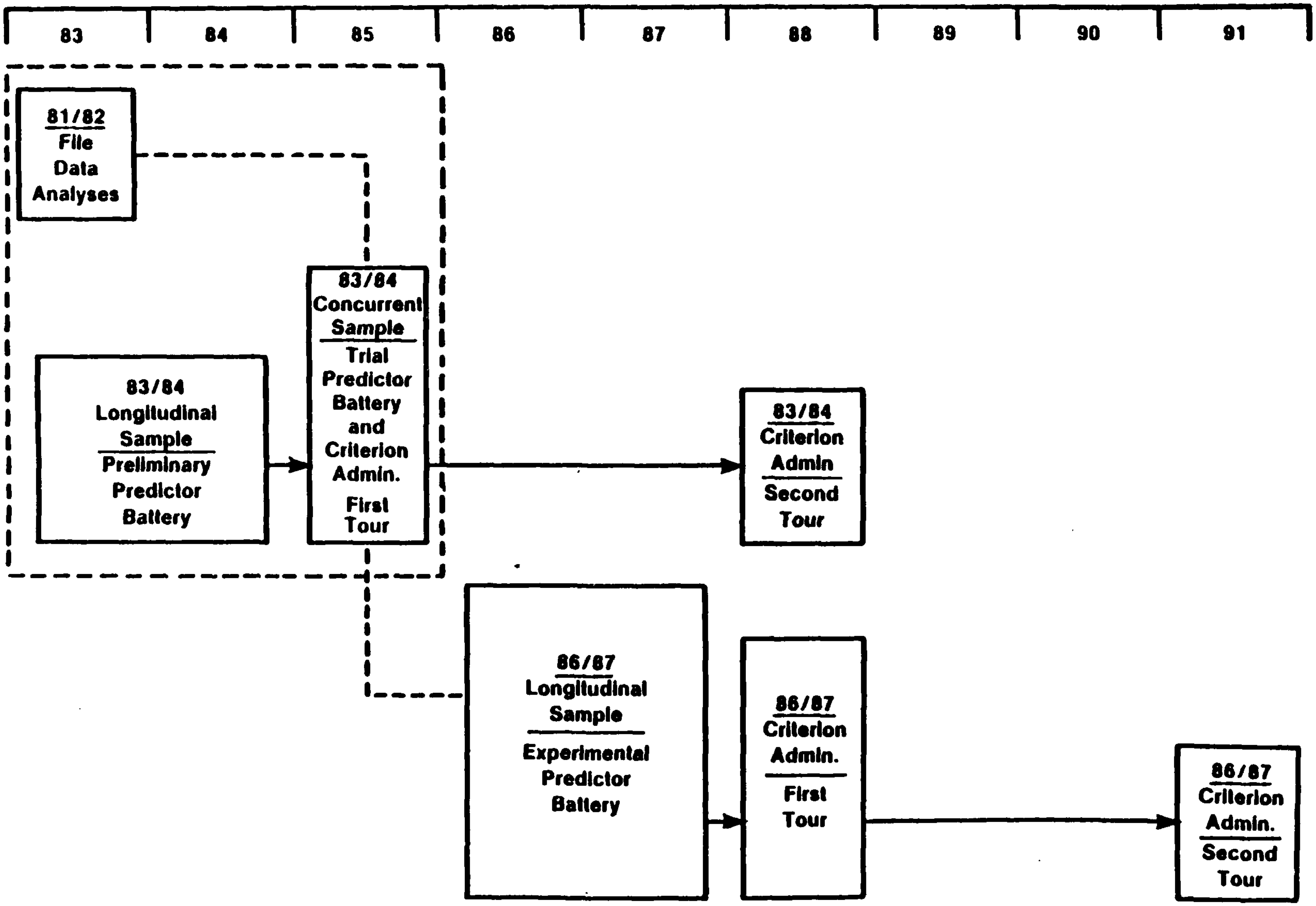
Governance Advisory Group

MG W.G. O'Leksy
CHAIRMAN

SCIENTIFIC ADVISORS	U.S. ARMY ADVISORS	INTERSERVICE ADVISORS
DR. P. BOBKO	MG J.B. ALLEN, JR.	DR. N.K. EATON (Chair)
DR. T. COOK	MG J.H. CORNS	DR. N. GUINN (AF)
DR. M. HAKEL (Chair)	MG T.J.P. JONES	DR. W.S. SELLMAN (DoD)
DR. L. HUMPHREYS	BG P. MALLORY	DR. M.F. WISKOFF (N)
DR. L. JOHNSON		
DR. R. LINN		
DR. M. TENOPYR		
DR. J. UHLANER		

FIGURE 3

THE RESEARCH FLOW


83
84
85
86
87
88
89
90
91
81/82 File Data Analyses
83/84 Longitudinal Sample Preliminary Predictor Battery
83/84 Concurrent Sample Trial Predictor Battery and Criterion Admin. First Tour
83/84 Criterion Admin Second Tour
86/87 Longitudinal Sample Experimental Predictor Battery
86/87 Criterion Admin. First Tour
86/87 Criterion Admin. Second Tour

The three waves of new data collected by the project consist of (1) a longitudinal sample called the preliminary battery sample which is composed of approximately 2,000 recruits in each of 4 MOS, (2) a major concurrent validation sample composed of 400-600 incumbents in each of 19 MOS, and (3) an even larger longitudinal validation sample composed of over 40,000 recruits taken from 21 MOS. Besides providing different kinds of validity information, the three samples were intended to provide the opportunity for multiple revisions of the new predictor battery. The preliminary battery sample was assessed with a four hour battery of carefully selected off-the-shelf tests to provide a set of marker variables for the project developed tests. Approximately one fifth of this sample became a part of the concurrent validation sample, which was the first time the full array of project developed tests and performance measures were administered together. Each job incumbent in the cross validation sample was assessed eight hours each day for two days. The longitudinal validation builds upon the concurrent findings and is designed to yield a sample of 400-600 per MOS after the decay rates for the MOS cohorts have their effect. To produce a sample of 10,000 incumbents at the time of job performance assessment, approximately 45,000 new recruits are being tested on the predictor battery.

The reenlistees from both the concurrent sample (83/84) cohort and from the longitudinal sample (86/87 cohort) will be followed into their second tour and assessed with another array of job performance measures. During the second tour the job tasks require a higher level of skill and the supervisor/leadership component becomes much more prominent.

The Sample

Since the ultimate goal is a new or modified classification system for making job assignments, it was necessary to sample jobs from the population of jobs for which the system would be used. We could not collect data on all 275 jobs. It was also necessary to obtain sufficient sample sizes for each job such that estimates would be reasonably stable. Additional considerations were a desire on the part of the Army to oversample the combat specialties and a collective desire to oversample MOS with higher proportions of women and/or minority groups.

I suppose the textbook procedure would be to calculate the required sample size via the appropriate power functions after deciding upon alpha, beta, and the critical effect size and then select a random sample of jobs from the population stratified by combat vs. support, racial composition, and gender composition. Unfortunately, it is not so easy to calculate power functions for multivariate effects, particularly when the dimensionality of the prediction equation is not yet known. There was also the matter of costs and the difficulty of finding the people when the time came for their assessment. Such considerations present a series of trade offs between the number of jobs and the number of people per job.

The sampling of jobs was made even less straightforward because even though the Army is a large organization, fewer than half the jobs include enough incumbents to make empirical validation feasible. Also, some jobs were in the process of being phased out while others were being changed. To make matters worse, job incumbents for some jobs are so scattered world-wide that obtaining reasonable sample sizes from them would be prohibitively expensive. Finally, the management of the organization cares more about accurate predictions of performance in some jobs than in others. These organizational dynamics are seldom mentioned as part of the estimation problem.

What finally transpired was that all MOS with more than 300 accessions per year were cluster analyzed on the basis of their judged similarity of task content. Then, given the considerations mentioned above, jobs were sampled "haphazardly" so as to represent the major content clusters, to oversample the relevant strata, and to keep the data collection costs within the project's budget limitations. The final degree of representativeness was determined not by adherence to true random sampling but by an intensive review of the entire process, procedures, and sampling outcome by management and several expert review panels. As a result, 19 jobs were chosen for study with projected sample sizes of 500-600 people each. The MOS sample is shown in Table 2.

Predictor Development

The standard operating procedure for predictor development in personnel selection research is to do a job analysis first. On the basis of a job analysis, the knowledges, skills and abilities required for successful performance are inferred, and an additional judgment is then made about which KSA's are trainable and which must be selected for. We didn't precisely do that in Project A (and have been criticized for it).

Instead, the strategy was to identify a universe of potential predictor constructs appropriate for the population of enlisted MOS, sample representatively from it, construct tests for each construct sampled, and refine and improve the measures through a long series of pilot and field tests. The intent was to develop a predictor battery that was maximally useful for an entire population of jobs and not to tailor-make them for the specific jobs in the sample. The loss in specific prediction accuracy for the jobs in the sample (if any) should be compensated for by the gain in coverage for all other jobs in the population.

The long process of predictor development is represented in Figure 4.

It began with an exhaustive search of the entire personnel selection literature. Teams were created for cognitive abilities, perceptual and psychomotor abilities, and non cognitive characteristics such as personality, interest, and biographical history. Every available automated and manual technique was used in the search and an initial list of several hundred variables was compiled. The list went through several waves of expert review and eventually came down to a list of 53 potentially useful predictor constructs. They are listed in Table 3.

Table 2. **PROJECT A MOS**

BATCH A	
11B	Infantryman
13B	Cannon Crewman
19E	Tank Crewman
31C	Radio TT Oper.
63B	Vehicle & Generator Mech.
64C	Motor Transport Oper.
71L	Admin. Specialist
91A	Medical Care Specialist
95B	Military Police

BATCH Z	
12B	Combat Engineer
16S	MANPADS Crewman
27E	Tow/Dragon Rpr.
51B	Carpentry/Masonry Spec.
54E	Chemical Operations Spec.
55B	Ammunition Spec.
67N	Utility Helicopter Rpr.
76W	Petroleum Supply Spec.
76Y	Unit Supply Spec.
94B	Food Service Spec.

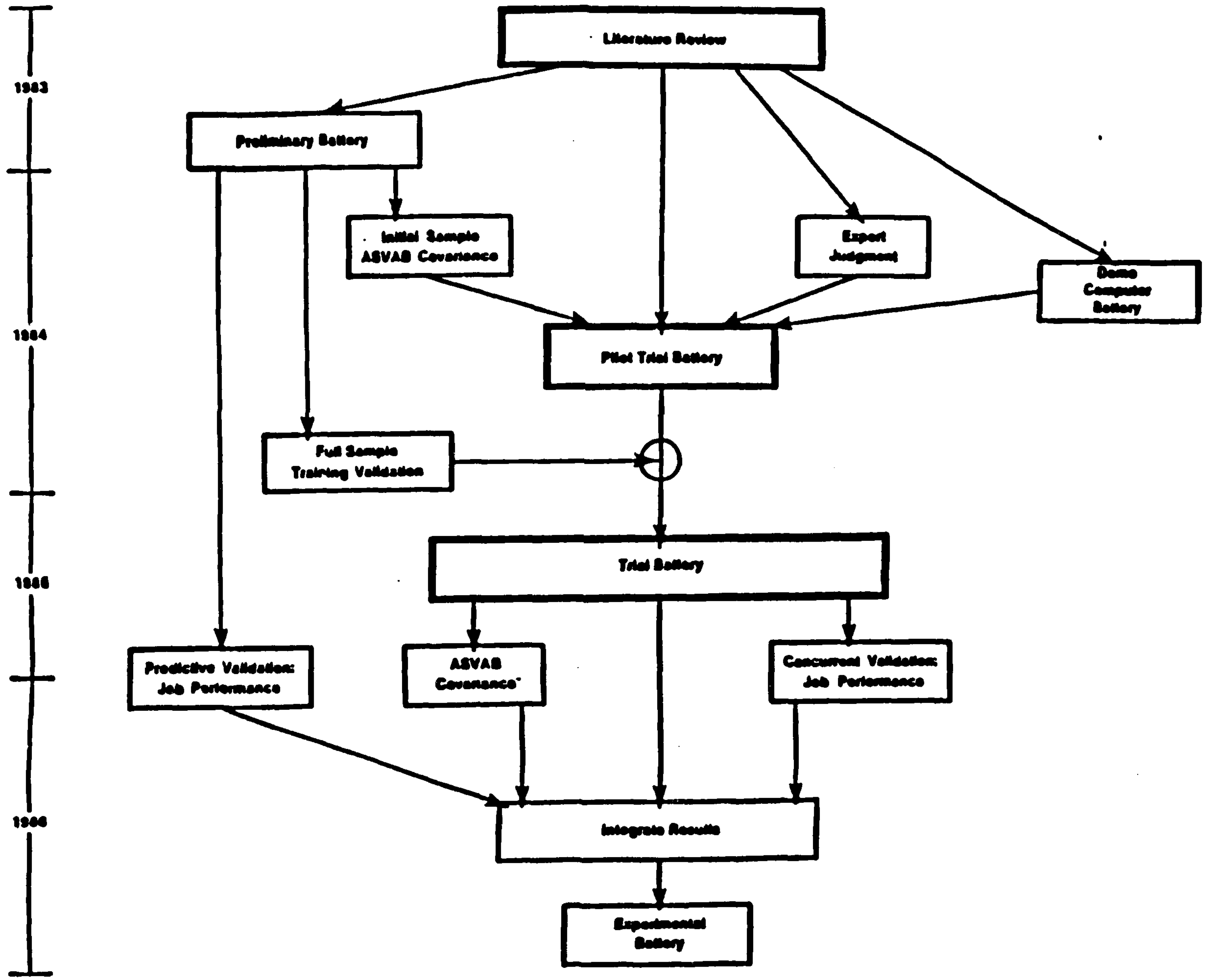


FIGURE 4. Flow Chart of Predictor Measure Development

CONSTRUCTS	CLUSTERS	FACTORS
1. Verbal Comprehension 5. Reading Comprehension 16. Ideational Fluency 18. Analogical Reasoning 21. Omnibus Intelligence/Aptitude 22. Word Fluency	A. Verbal Ability/ General Intelligence	COGNITIVE ABILITIES
4. Word Problems 8. Inductive Reasoning: Concept Formation 10. Deductive Logic	B. Reasoning	
2. Numerical Computation 3. Use of Formula/Number Problems	C. Number Ability	
12. Perceptual Speed and Accuracy	H. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory 17. Follow Directions	J. Memory	
19. Figural Reasoning 23. Verbal and Figural Closure	F. Closure	
6. Two-dimensional Mental Rotation 7. Three-dimensional Mental Rotation 9. Spatial Visualization 11. Field Dependence (Negative) 15. Place Memory (Visual Memory) 20. Spatial Scanning	E. Visualization/Spatial	VISUALIZATION/ SPATIAL
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental Information Processing	INFORMATION PROCESSING
13. Mechanical Comprehension	L. Mechanical Comprehension	MECHANICAL
48. Realistic Interests 51. Artistic Interests (Negative)	M. Realistic vs. Artistic Interests	
28. Control Precision 29. Rate Control 32. Arm-hand Steadiness 34. Aiming	I. Steadiness/Precision	PSYCHOMOTOR
27. Multilimb Coordination 35. Speed of Arm Movement	D. Coordination	
30. Manual Dexterity 31. Finger Dexterity 33. Wrist-Finger Speed	K. Dexterity	
39. Sociability 52. Social Interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising Interests	R. Enterprising Interest	
36. Involvement in Athletics and Physical Conditioning 37. Energy Level	T. Athletic Abilities/Energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/Self-esteem	
40. Traditional Values 43. Conscientiousness 46. Non-delinquency 53. Conventional Interests	N. Traditional Values/Conventionality/Non-delinquency	MOTIVATION/ STABILITY
44. Locus of Control 47. Work Orientation	O. Work Orientation/Locus of Control	
38. Cooperativeness 45. Emotional Stability	P. Cooperation/Emotional Stability	

Table 3. Hierarchical Map of Predictor Space

A similar, but different, procedure was used to identify a population of performance factors - 72 in all. We then assembled a sample of 35 personnel selection experts and asked them to estimate the correlation between each predictor construct and each criterion factor, when that correlation was corrected for restriction of range and criterion unreliability. The resulting judgments could be analyzed for the inter judge agreement, rows and columns could be factor analyzed, and the results could be compared to analogous information from the empirical literature. Most importantly however, the exercise provided another substantial set of expert judgments about which predictor constructs should be the most useful. A hierarchical analysis of the predictor validity profiles is also shown in Table 3.

All the available information was then used to arrive at a final set of variables for which new measures would be constructed. This represented months of effort by lots of people to select the variables that will best supplement the ASVAB in predicting job performance across all MOS. What followed were many months more of instrument construction, several waves of pilot tests, and a series of major field tests. Included in these efforts were the development of a computerized battery of perceptual/psychomotor tests, the creation of the software, the design and construction of a special response pedestal permitting a variety of responses (e.g., one hand tracking, two hand coordination) and the acquisition of 108 portable computerized testing stations. After each data collection, revisions were made on the basis of item statistics and expert review. Finally on May 15, 1985 the predictor battery was deemed ready for concurrent validation. That battery, known as the Trial Battery (TB), is listed in Table 4.

Performance Measurement

The goals of training and job performance measurement in Project A were to define, or model, the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor.

Some additional specific goals were to: a) make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency, b) compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e. a multi-trait, multi-method approach), c) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and d) evaluate existing archival and administrative records as possible indicators of job performance.

Given these intentions, the criterion development effort focused on three major methods: hands-on job sample tests, multiple choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in the development of the rating methods.

Modeling Performance

The development efforts to be described were guided by a particular "theory" of performance. The basic outline is as follows.

Table 4

Summary of Predictor Measures Used in Concurrent Validation
(the Trial Battery)

COGNITIVE PAPER-AND-PENCIL TESTS

Test Name (Construct Name)	Number of Items
Reasoning Test (Induction-figural reasoning)	30
Orientation Test (Spatial orientation)	24
Map Test (Spatial orientation)	20
Object Rotation Test (Spatial Visualization - Rotation)	90
Assembling Objects Test (Spatial visualization - Rotation)	32
Maze Test (Spatial visualization - scanning)	24

COMPUTER-ADMINISTERED TESTS

Name	Number of Items
Simple Reaction Time (Processing efficiency)	15
Choice Reaction Time (Processing efficiency)	30
Memory Test (Short-term memory)	36
Target Tracking Test #1 (Psychomotor precision)	18
Target Shoot Test (Psychmotor precision)	30
Perceptual Speed and Accuracy Test (Perceptual speed & accuracy)	36
Identification Test (Perceptual speed and accuracy)	36
Target Tracking Test #2 (Two hand coordination)	18
Number Memory Test (Number operations)	28
Cannon Shoot Test (Movement judgment)	36

NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES

Inventory Name and Subscale Name	Number of Items
Assessment of Background and Life Experiences (ABLE) Inventory	209
Adjustment	
Dependability	
Achievement	
Physical Condition	
Leadership	
Locus of Control	
Agreeableness/Likeability	
Army Vocational Interest Career Examination (AVOICE)	176
Realistic Interests	
Conventional Interests	
Social Interests	
Enterprising Interests	
Artistic Interests	

First, job perfoımance really is multi-dimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization.

Two General Factors

For the population of entry level enlisted positions we postulated that there are two major types of job performance components. The first is composed of components that are specific to a particular job. That is, measures of such components would reflect specific technical competence or specific job behaviors that are not required for other jobs. The second kind of performance factor includes components that are defined and measured in the same way for every job. These are referred to as Army-wide criterion factors.

For the job specific components we anticipated that there would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills and which would be reflected by different task content.

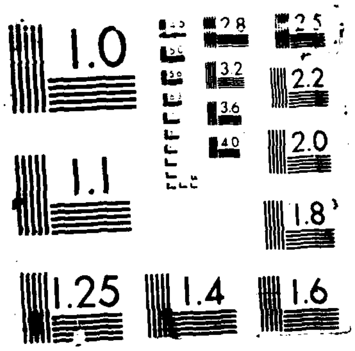
The Army-wide concept incorporates the basic notion that total performance is much more than task or technical proficiency. It might include such things as contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity.

In sum, the working model of total performance with which the project began viewed performance as multi-dimensional within the two broad categories of factors or constructs. The job analysis and criterion construction methods were designed to "discover" the content of these factors via an exhaustive description of the total performance domain, several iterations of data collection, and the use of multiple methods for identifying basic performance factors.

Factors vs. a Composite

Saying that performance is multi-dimensional does not preclude using just one index of an individual's contributions to make a specific personnel decision (e.g., select/not select, promote/not promote). As argued by Schmidt and Kaplan (1971) some years ago, it seems quite reasonable for the organization to scale the importance of each major performance factor relative to a particular personnel decision that must be made and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision. That is, the way in which performance information is weighted and combined is a value judgment on the organization's part. The determination of the specific combinational rules (e.g., simple sum, weighted sum, non linear combination) that best reflect what the organization is trying to accomplish is a matter for research.

AD-A198 274 IMPROVING THE SELECTION CLASSIFICATION AND UTILIZATION 3/5
OF ARMY ENLISTED P (U) HUMAN RESOURCES RESEARCH
ORGANIZATION ALEXANDRIA VA J P CAMPBELL MAY 88
UNCLASSIFIED HUMRRO-IR-PRD-87-12-SUPPL F/G 5/9 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

A Structural Model

If performance is characterized in the above manner, then a more formal way to model performance is to think in terms of its latent structure, postulate what that might be and then resort to a confirmatory analysis. Unfortunately, it is true that we simply know a lot more about predictor constructs than we do about job performance constructs. There are volumes of research on the former, and almost none on the latter. For personnel psychologists it is almost second nature to talk about predictors in terms of theories and constructs. However, on the performance side the textbooks are virtually silent. Only a few people have even raised the issue (e.g., Dunnette, 1963; Wallace, 1965).

Unit vs. Individual Performance

Finally, people do not usually work alone. Individuals are members of work groups or units and it is the unit's performance that frequently is the most central concern of the organization. However, determining the individual's contribution to the unit's performance score is not a simple problem. Further, variation in unit performance is most likely a function of a number of factors besides the "true" level of performance of each individual.

For two major reasons, Project A has not incorporated unit effectiveness in its model of performance. First, the project is focused on the development of a new selection/classification system for entry level personnel and is concerned with improving personnel decisions about individuals and not units. The task is to maximize the average payoff per individual selected.

The second major reason is the prohibitive cost. It simply was not possible to develop reliable and valid field exercises for assessing unit performance in a representative sample of jobs within a reasonable time frame. In isolated instances it might be possible to take advantage of regularly scheduled exercises or use existing performance records that a particular unit (e.g., maintenance depot) might keep. However, it proved not possible to obtain such data in any systematic way. Even if it could be done, it would not be easy to establish the correspondence between individual performance and unit effectiveness.

What we have chosen to do is to try to identify the factors, or means, by which individuals contribute to unit performance and to assess individual performance on those factors via rating methods. We also have a certain amount of information on situational and unit characteristics and are attempting to determine how much of the variance in individual performance is accounted for by those characteristics.

Criterion Development

Actual criterion development proceeded from two basic types of information. First, all available task descriptions were used to generate a population of job tasks for each MOS. The principal sources of task description are the Army's periodic job description surveys, which use questionnaire checklists of several hundred task statements to survey job incumbents about the frequency with which they perform each task, and the Soldier's Manual for each job which is a complete specification by management of what the task content of the job is supposed to be. The two sources describe tasks at a somewhat different level of generality with the occupational survey items being much more specific in nature.

Unfortunately no textbook or available technology tells us what the specifications of a task description should be for different purposes. We opted for statements which described a complete operation which had a recognizeable beginning and end, and which was relatively independent of other tasks. That is, it is possible to perform Task A without performing Task B. After much editing, revising, and a formal review by a panel of subject matter experts, a population of 130-180 tasks was enumerated for each MOS.

An additional series of expert judgments was then used to scale the relative difficulty and importance of each task and to cluster tasks on the basis of content similarity. Sampling tasks for measurement was accomplished via a kind of Delphi procedure. That is, each member of a team of task selectors was asked to select 30 tasks from the population of tasks such that the selected tasks were representative of task content, were important, and represented a range of difficulty. The individual judge's choices were then regressed on the task characteristics and both the choices and the captured "policy" of each person were fed back to the group members, who each revised their choices as they saw fit. Typically, convergence was achieved quickly and the final selection was by consensus. The consensus of the task selection panel was then thoroughly reviewed by the Army command responsible for that particular job. This last review was the "final" word on the representativeness of task samples and produced a sample of 30 tasks for each job.

Standardized job samples, the paper-and-pencil job knowledge tests, and numerical ratings scales were then constructed to assess knowledge and proficiency on these tasks. Each measure went through multiple rounds of pilot testing and revision. The job sample tests were fairly elaborate and were composed of multiple test stations sometimes spread over a football field size area. Each task to be tested was broken down into several steps each of which was scored pass/fail.

The second procedure used to describe job content was the critical incident method. Panels of NCO's and officers generated thousands of critical incidents of effective and ineffective performance. There were two basic formats for the critical incident workshops. One asked participants to generate incidents that potentially could occur in any job. The second type focused on incidents that were specific to the content of the particular job under consideration. The behaviorally anchored rating scale procedure was used to construct rating scales for performance factors specific to a particular job (MOS Specific BARS) and performance factors that were defined in the same way and relevant for all jobs (Army-wide BARS).

The critical incident procedure was also used with workshops of combat veterans to develop rating scales of "expected" combat effectiveness.

Since one major objective was to determine the relationships between training performance and job performance and their differential predictability, if any, a comprehensive training achievement test has been constructed for each MOS by carefully matching the content of the program of instruction (POI) with the content of the population of job tasks, and writing items to represent each segment of the match. We were most interested in task content which is taught, and also performed on the job, versus tasks which were performed on the job but not part of the POI. Scores on this latter category of items (when given to trainees) would be a measure of incidental learning. The correlation of direct learning and incidental learning with job performance, both when initial ability is controlled and where it is not, is of considerable interest.

The final entry in the array of criterion measures was produced by a concerted effort to get what we could from the files or archival records. Potentially at least, there are numerous performance indicators lurking in existing computer records and personnel files. We began by enumerating all possibilities from three major sources of such records.

<u>The Enlisted Master File</u> (EMF) - a central computer record of selected personnel actions.

<u>The Enlisted Military Personnel File</u> (OMPF) - which is the permanent historical record of an individual's military service kept on microfische at a central location.

<u>Military Personnel Records Jacket</u> (MPRJ) - or more commonly known as the 201 file which is the personnel folder that follows the individual.

We systematically compared these three sources using a sample of 750 people and a standardized information recording form. The 201 file looked the most promising in terms of recency and completeness, but of course it is by far the most expensive to search. (The textbooks never mention these cost benefit questions). As a consequency, everyone crossed their fingers and we collected eight archival performance indicators via a self report questionnaire. That is, people were asked what was in their personnel file as regards letters of commendation, disciplinary actions,, etc.. Field tests on a sample of 500 people showed considerable agreement between self report and archival records. Almost all disagreements were in the direction of more frequent self reports, for <u>both</u> positive and negative things. Further followup questionnaires and interviews suggested that self report may be the more accurate. Anyway, we used them and their distributions and correlations seemed quite reasonable. The self report items were combined into 4 indicators that were actually used as criterion measures.

The complete array of performance measures in the form in which they survived a large scale field study of N = 150/MOS for nine MOS is shown in Table 5.

These are the measures which were administered to the concurrent sample of 400-600 people in each of the 19 MOS. The distinction between the Batch A (9 MOS) and Batch Z (10 MOS) is that not all criterion measures were developed for each job. Budget constraints dictated that the job specific measures could only be developed for a limited number of jobs (i.e. Batch A).

Table 5

Summary of Criterion Measures Used in Concurrent Validation Samples[1]

Performance Measures Common to Batch A and Batch Z MOS (Jobs)

1. Ten behaviorally anchored rating scales designed to measure factors of non job specific performance (e.g., Giving peer leadership and support, maintaining equipment, self discipline).

2. Single scale rating of overall job performance.

3. Single scale rating of NCO (Non Commissioned Officer) potential.

4. Ratings of performance on 13 representative "common" tasks. The Army specifies a series of common tasks (e.g., several first aid tasks) that everyone should be able to perform.

5. Paper-and-pencil Test of Training Achievement developed for each of the 19 MOS (130-210) items each).

6. A 77 item summated rating scale for the assessment of expected combat performance.

7. Five performance indicators from administrative records. The first three are obtained via self report and the last two from computerized records.

    - Total number of awards and letters of commendation.
    - Physical fitness qualification.
    - Number of disciplinary infractions.
    - Rifle marksmanship qualification score.
    - Promotion rate (in deviation units).

Performance Measures for Batch A Only

8. Job-sample (hands-on) test of MOS-specific task proficiency.

    - Individual is tested on each of 15 major job tasks.

9. Paper-and-pencil job knowledge tests designed to measure task specific job knowledge.

    - Individual is scored on 150-200 multiple choice items representing representing 30 major job tasks. Fifteen of the tasks were were also measured hands-on.

10. Rating scale measures of specific task performance on the 15 tasks also measured with the knowledge tests and the hands-on measures.

11. MOS-specific behaviorally anchored ratings scales. From 6 to 10 BARS scales were develcped for each MOS to represent the major factors that constituted job specific technical and task proficiency.

Auxiliary Measures Included in Criterion Battery

- A Job History Questionnaire which asks for information about frequency and recency of performance of the MOS-specific tasks.

- Work Environment Description Questionnaire - a 141 item questionnaire assessing situational/environmental characteristics, leadership climate, and reward preferences.

[1] All rating measures were obtained from approximately 2 supervisors and 3 peers for each ratee.

Data Preparation

Between July 1 and December 1 of 1985 the predictor and criterion batteries were administered to 9430 job incumbents. Four hours were devoted to the predictor tests and 12 hours to the criterion measures. All this presented a large data collection task. Eight person teams supported by 4 to 5 Army personnel visited each of 15 different Army posts for several weeks at a time. The logistics and complexity and magnitude of the data collection were expected but no one was prepared for the subsequent massiveness of the data matching and editing task. Considerable effort was devoted to training the data collection teams, standardizing testing conditions, keeping logs, and performing data checks each day. Hundreds and hundreds of hours have been dedicated to identifying mismatched data and finding missing data. However, we still have complete records for eleven people we can't identify. The textbooks wisely don't mention how noxious all this is. It would keep people from entering the field.

While the amount of missing data is not large when considered instrument by instrument, if complete data one very instrument were demanded for each individual, the total sample size would shrink by 80%. The majority of people are missing at least one test item or one rating scale. However, most of the people missing data were not missing very much and we developed a series of decision rules for dropping cases with too many elements missing and imputing data, by various methods, for the remainder. There certainly are no _a priori_ rules for designating how much is too much or for identifying the appropriate method of imputation. We again had to rely on a consensus of expert judgment helped along by staring at a lot of cumulative distributions and computing covariance matrices before and after inputing data.

As of now we have clean data and reasonable sample sizes. All that remains is to find what we set out to look for in the first place.

Results From the Concurrent Validation Sample

If all the rating scales are used separately, the MOS-specific measures are aggregated at the task or instructional module level, and the major predictor subscales are used, there are approximately 200 criterion scores and 60-70 predictor scores on each individual.

At this point a classic argument arises between the empirical keying/ "let's look for all the specific variance we can" types and the individuals who want to reduce collinearity as much as possible and deal at the construct level. We have tried for more of the latter than the former for a number of reasons. One reason is that we would like the project to produce as many generalizeable truths as possible. Another stems from the dilemma between accuracy of prediction and accuracy of estimation, or the cross validation problem. That is, the more a prediction equation maximizes the accuracy of prediction in the sample, the more error it introduces into the estimation of the degree of accuracy in the population.

Project A is faced with the task of estimating several kinds of differential validity. It is reasonable to ask at the outset whether it is even possible, for a system of any multivariate complexity, to detect reasonable amounts of differential prediction with reasonable amounts of statistical power. The fewer parameters one must estimate, the greater the chances of being able to do that, which is a primary reason for examining the latent structure of predictors and criteria as carefully as possible.

Since we can draw a fairly reasonable picture of the population covariance matrices for both predictors and criteria and thus provide a better starting point for Monte Carlo studies, one major research question we hope to answer is whether it is <u>ever</u> possible to estimate the parameters necessary for building a true classification algorithm. If it can't be done with a sample of 20 jobs and 500 cases per job then perhaps the textbook discussions of the classification problem are a bit academic.

The Road to Constructs

For both predictors and criteria, the procedure for getting from the individual task or scale scores to factor or construct scores was similar; except for the degree to which the previous literature was of help. Many decades of research on the measurement of abilities, personality, and interests have provided a lot of information about the structure of individual differences. Similar help from the performance side is really not available except for a modest number of descriptive studies of specific occupations such as managers, nurses, police officers, fire fighters, and the elusive and seldom seen college professor. Unfortunately, we were operating in a different job population and knew only that paper-and-pencil measures and rating measures would produce a lot of so-called method variance.

Given this initial disparity, we used both expert judgment and factor analytic results from the field tests to formulate a model target. A picture of that model is shown in Figure 5.

I include this only to show one stage in the almost continuous process of bootstrapping ourselves toward a more final conceptual description of the predictor/criterion space.

These targets were then subjected to what might best be described as a "quasi" confirmatory analysis using the concurrent validation sample. For the predictor scales that meant using the target to specify the number of factors for a full sample solution (i.e. all MOS combined). The predictor constructs and their associated component scales are shown in Table 6.

For the within MOS criterion matrices we used confirmatory analyses and attempted to test alternative models. The alternative models were obtained by allowing the principal investigators to peek at the data, in the form of a series of principal component analyses, and to formulate a target matrix for a LISREL solution. Some clear alternative ideas emerged and these were compared in each MOS. After not too much cutting and fitting, we arrived at a single portrayal of the latent structure of performance that both fit the data

FIGURE 5

JOB PERFORMANCE — A PROPOSED STRUCTURAL MODEL

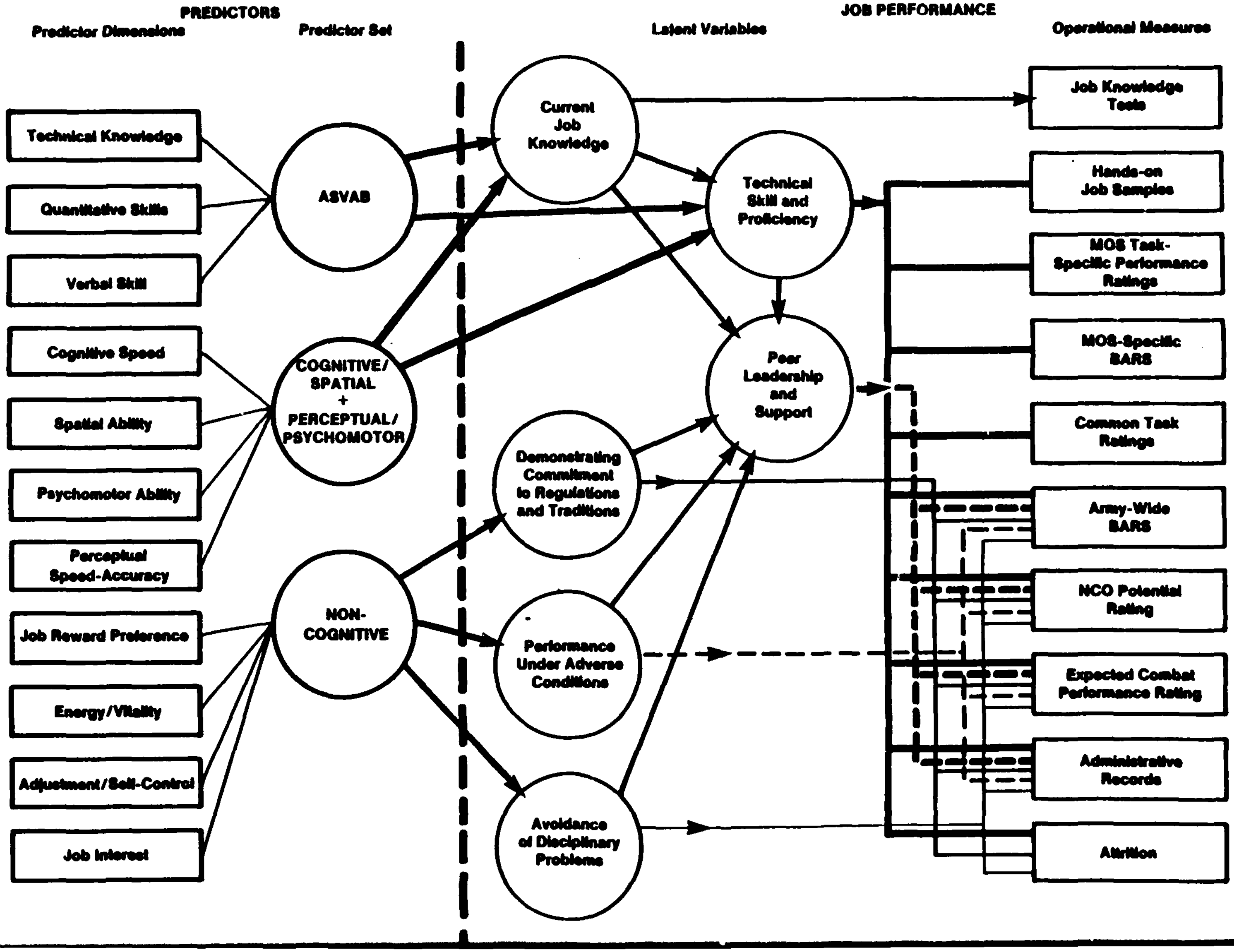


Table 6

Ability, temperament, andinterest factors identified via analysis of the concurrent validity data on 9430 job incumbents.

FROM PAPER-AND-PENCIL TESTS

Overall Spatial Factor
- Assembling Objects test
- Map test
- Maze test
- Object Rotation test
- Orientation test
- Figural Reasoning test

FROM COMPUTERIZED MEASURES

Psychomotor Factor
- Cannon Shoot test (Time score)
- Target Shoot test (Time to fire)
- Target Shoot test (Log distance)
- Target Tracking 1 (Log distance)
- Target Tracking 2 (Log distance)
- Target Tracking 2
- Short Term Memory test (Decision time)

Perceptual Speed/accuracy Factor
- Short Term Memory test (Percent correct)
- Perceptual Speed & Accuracy test (Decision time)
- Perceptual Speed & Accuracy test (Percent correct)
- Target Identification test (Decision time)
- Target Identification test (Percent correct)

Number Speed/accuracy Factor
- Number Memory test (Percent correct)
- Number Memory test (Initial decision time)
- Number Memory test (Mean operations decision time)
- Number Memory test (Final decision time)

General Reaction Speed Factor
- Choice Reaction Decision Time
- Simple Reaction Decision Time

General Reaction Accuracy Factor
- Choice Reaction Percent Correct
- Simple Reaction Percent Correct

FROM NON-COGNITIVE INVENTORIES

Achievement Factor
- Self-esteem scale
- Work Orientation scale
- Energy Level scale

Dependability Factor
- Conscientiousness scale
- Non-delinquency scale

Adjustment Factor
- Emotional Stability scale

Physical Condition Factor
- Physical Condition scale

Skilled Technician Interest Factor
- Clerical/Administrative
- Medical Services
- Leadership/Guidance
- Science/Chemical
- Data Processing
- Mathematics
- Electronic Communications

Structural/Machines Interest Factor
- Mechanics
- Heavy Construction
- Electronics
- Vehicle/Equipment Operator

Combat Related Interest Factor
- Combat
- Rugged Individualism
- Firearms Enthusiast

Audiovisual Arts Interest Factor
- Drafting
- Audiographics
- Aesthetics

Food Service Interest Factor
- Food Service Professional
- Food Service Employee

Protective Services Interest Factor
- Law Enforcement
- Fire Protection

NOTE: The tests and inventory scales from the trial battery which were used to form simple sum factor scores are listed under each factor title.

in each job and seemed to make good sense. Obviously, the confirmatory analysis was not used in a strictly confirmatory way. This structure of job performance is portrayed in Table 7.

The model best confirmed by LISREL specified five "substantive" and two methods factors which we labeled the "ratings" factor and the "test" factor. The ratings factor was specified to be the first orthogonal component taken from all the rating scales. The test factor is the first orthogonal component taken from the paper-and-pencil knowledge tests. Given this constraint, five substantive factors were extracted. The first two are based on the knowledge tests and the job sample measures. We have called these the core technical performance factor and the general (not so core) task performance factor. The technical factor reflects content which is central and largely specific to the MOS. The second factor encompasses content that tends to be common across several jobs and is less central to the core performance objectives. For this job population a significant part of the factor is represented in the common tasks, such as first aid, basic navigation, use of communication equipment, etc.. However, I expect it should be possible to make this distinction for virtually any job.

The remaining factors are based on the ratings, primarily those developed by the critical incident method, and the administrative/personnel records that were collected via self-report. Factor three encompassed the most scales and was the clearest in terms of its loading but the most heterogeneous appearing in terms of content. It appears to be a general effort and performance, performance under adverse conditions, peer leadership factor. In a spirit of wishful thinking we had originally hoped to separate some of these elements, but either the lack of a distinct latent structure or the fallibility of the measures prevented it. More about this factor in a minute. Factor four is much more homogeneous and reflects the rating scales having to do with personal discipline and avoidance of trouble and the number of negative personnel outcomes people reported. Factor five is fairly narrow in content and shows very clear loadings for ratings of military bearing and the physical fitness score that is part of everyone's personnel record. This factor made me wonder what would happen if we looked for an appearance factor in non-military occupations. While the military has good reasons for paying for performance on such a factor, one of the most disheartening findings in all of psychology is the degree to which people associate all kinds of good things with good looks.

In general, this solution fits the data from all MOS, seems reasonable and appropriate to Army management, and is not too far from our hypothesized structure, although we hoped to split factors two and three into a few more pieces. This picture also seems like a useful starting point for analyses of additional non-military jobs.

Given these two pictures of the predictor domain and the performance space, we have barely begun exploring questions of differential validity across criterion components, differential validity across jobs, differential validity across subgroups of people, and overall classification efficiency under a variety of constraints. The bulk of the analyses are yet to be done. However, let me close with a few tidbits that I hope will keep you coming back year after painstaking year.

Table 7

Performance factors representing the common latent structure across all jobs in Project A sample.

1) **Task Proficiency: MOS (Job) specific core technical skills**: The proficiency with which the individual performs the tasks which are "central" to his or her job (MOS). The tasks represent the core of the job and they are its primary definers from job to job.

- The subscales representing core content in both the knowledge tests and the job sample tests that loaded on this factor were summed within method, standardized, and then added together for a total factor score. The factor score does not include any rating measures.

2) **Task Proficiency: General or common skills**: In addition to the core technical content specific to an MOS, individuals in every MOS responsible for being able to perform a variety of general or common tasks --e.g., use of basic weapons, first aid, etc.. This factor represents proficiency on these general tasks.

- The same procedure (as for factor one) was used to sum the general task scales, standardized within methods, and add the two standardized scores.

3) **Peer Leadership, Effort, and Self Development**: Reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient.

- Five scales from the Army-wide BARS rating form (gen. tech. performance, peer leadership, demonstrated effort, self development, gen. maintenance), the expected combat performance scales, the job specific BARS scales, and the total number of commendations and awards received by the individual were summed for this factor.

4) **Maintaining Personal Discipline**: Reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self control, demonstrates responsibility in day to day behavior, and does not create disciplinary problems.

- Scores on this factor are composed of three Army-wide BARS scales (adherence to traditions and regulations, exercising self control, demonstrating integrity), a subscale from the combat rating pertaining to avoidance of trouble, and two indices from the administrative records (number of disciplinary actions and promotion rate).

5) **Military Bearing/Appearance**: Represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

- Factor scores are the sum of the physical fitness qualification score from the individual's personnel record and the "military bearing and appearance " rating scale.

NOTE: The criterion measures that comprise each factor are as indicated.

Differential Prediction Across Criterion Components

First, the different criterion components are not predicted by the same things. We have begun to look at cognitive abilities vs. spatial/perceptual/psychomotor abilities vs. temperament/interests.

Table 8 shows the multiple correlation of the components in these domains (corrected for shrinkage and for restriction in range, but not for unreliability) with the five criterion factors.

The entries in the table represent the average across all MOS. The level of validity of ASVAB for the first two factors is about the same as, or higher than, that usually observed when ASVAB is correlated with training criteria. ASVAB does predict job performance. For the third factor the validity of the cognitive tests drops, but is still substantial, and the validity of the non cognitive inventories increases. This reversal becomes even more distinct for factors 4 and 5. Notice that the interest scales are also a reasonable good predictor of task performance and do not predict factors 3, 4, and 5 as well as the temperament scales. The mixed nature of factor three is interesting and along with the confounding of method variance between the first two and the last three factors, it invites a consideration of residual scores and that is what I would like to turn to next.

Another Look at Halo Error

The "ratings" and "test" methods factors were scored via regression methods to make them orthogonal to the other variables. The ratings methods factor was then partialled from the criterion construct scores that used ratings (i.e. factors 3, 4, and 5) and the "test" methods factor was partialled for each of the first two factors (i.e. those based in part on paper-and-pencil test scores). As a result, there are now 10 criterion scores -- 5 "raw" factors scores and 5 residual factor scores.

The mutliple correlation (corrected for shrinkage) of the different predictor groups with these ten scores is shown in Table 9.

For us at least, one of the most interesting aspects of the table is a comparison of the factor 3 raw score with the residualized factor 3. As compared to the correlations with the raw score the correlations of the cognitive measures with the residual go up substantially and the correlations with the temperament composite go down slightly. The correlation of the interest composite with factor 3 also goes up when the ratings method factor is paralleled out. In general, interest in task content is more closely associated with task performance than with the more volitional nature of factors 3, 4, and 5. These differences are not nearly so pronounced for the other two factors that involve ratings. We thinkg this is because factor three includes the scales that in fact asked raters to assess the technical performance of the ratee.

Table 8

Multiple correlations[1] of five independent predictor composites with each of five job performance criterion factors.

CRITERION FACTORS	ASVAB[2] Composite k = 4	Spatial Abilities Composite k = 1	Perceptual/ Psychomotor Abilities Composite (Computerized) k = 5	Temperament Scales and Bio data Composite ABLE k = 4	Interest Scales Composite Avoice k = 6
1. Core Technical Proficiency	61	54	49	24	33
2. General Task Proficiency	67	64	56	25	37
3. Effort, Leadership, Self Development	35	28	27	34	26
4. Personal Discipline	19	16	14	32	15
5. Physical Fitness & Military Appearance	21	11	11	37	12

PREDICTORS

NOTE: Entries in the table are averaged across the 9 jobs (MOS) in Batch A. N = 4400.

[1] Multiple R's are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2] k = the number of predictor scores in the composite.

Table 9

Multiple correlations[1] of five independent predictor composites with each of five job performance criterion factors and five residual scores formed by partialing a paper-and-pencil methods factor from (1) and (2) and a ratings methods factor from (3), (4), and (5).

CRITERION FACTORS		ASVAB[2] Composite k = 4	Spatial Abilities Composite k = 1	Perceptual/ Psychmotor Abilities Composite (Computerized) k = 5	Temperament Scales and Bio data Composite ABLE k = 4	Interest Scales Composite AVOICE k = 6
1	Core Technical Proficiency	61	54	49	24	33
1(r)	Core Technical Proficiency (Residual)	45	38	34	19	24
2	General Task Proficiency	67	64	56	25	37
2r	General Task Proficiency (Residual)	51	49	42	22	31
3	Effort, Leadership, Self Development	35	28	27	34	26
3r	Effort, Leadership, Self Development (Residual)	47	42	38	31	33
4	Personal Discipline	19	16	14	32	15
4r	Personal Discipline (Residual)	21	18	15	28	17
5	Physical Fitness & Military Appearance	21	11	11	37	12
5r	Physical Fitness & Appearance (Residual)	21	11	14	35	15

PREDICTORS

NOTE: Entries in the table are averaged across the 9 jobs (MOS) in Batch A. N = 4400.

[1] Multiple R's are adjusted for shrinkage and corrected for restriction in range, but not corrected for criterion unreliability.

[2] k = the number of predictor scores in the composite.

It is tempting to infer that raters are in fact influenced by the actual task competence of raters but that they also reflect differences in what might be termed dispositional or volitional behaviors of the kind predicted by personality/interest measures. Does the individual work hard, help others when they need it, keep going under adverse conditions, etc.? In our framework, these are both important components of performance and they are predicted by different things, but assessment via ratings cannot separate them very well. Perhaps it is also understandable why raters would have a difficult time separating them. It would require almost a mental partial correlation to do so.

Things Not Mentioned

There is not time to go into two additional parts of Project A that are underway but not concluded. I will just mention them.

Criterion Composites

First, we are currently carrying out a series of scaling studies in which supervisors and managers within each job (MOS) are being asked to judge the relative importance of each criterion component for overall performance when selection is the objective. Different methods and different situational contexts have been explored. So far we know that judges can use any one of several methods reliably, and that the patterns of weights also seem to differ across MOS. Whether weighting will make any difference in the end is another matter.

Performance Utility

A second ongoing effort is an attempt to scale the utility of job by performance level combinations. As I think Cronbach and Gleser demonstrated a number of years ago, to the extent that such differential utilities do exist and can be assessed, the potential payoff from classification is increased over and above that produced by differential prediction itself. This has not been an easy road to travel. The economists yell about marginal vs. average utility, the context for the judgment does make a difference, the appropriate metric is inclear since expressing utility in dollar terms does not seem appropriate for an organization such as the Army, and sooner or later top management must come face to face with its value judgments. All these difficulties aside, the goals are to see if it can be done and to determine, via Monte Carlo procedures, how much utility differences would influence personnel assignments when held against other constraints on classification (e.g., labor supply and demand).

In Conclusion

This then is an outline of Project A. In the next one to three years we hope to learn much more about the nature of job performance, the predictability of different aspects of performance, differential prediction across jobs, differential prediction across subgroups with performance components and within predictor domains, and something about the limits that the structure of individual differences places on selection and classification decisions. Stay tuned.

REFERENCES

Dunnette, M. D. (1963). A modified model for selection research. Journal of Applied Psychology. 47, 317-323.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. Personnel Psychology. 24, 419-434.

Wallace, S. R. (1965). Criteria for what? American Psychologist. 20, 411-418.

COMPARABILITY OF HANDS-ON AND KNOWLEDGE TESTS ACROSS NINE MILITARY JOBS

Patrick Ford and Charlotte H. Campbell
Human Resources Research Organization

Daniel B. Felker and Dorothy C. Edwards
American Institutes for Research

Michael G. Rumsey
U.S. Army Research Institute

Presented on symposium,
"Multiple Criteria and Multiple Jobs: Will One Model Fit All?"

At the Annual Convention of the
American Psychological Association

Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Comparability of Hands-On and Knowledge Tests Across Nine Military Jobs

Nine jobs, or military occupational specialties (MOS), were covered intensively in the Project A concurrent validation. The intensive coverage included MOS-specific rating scales and written tests and hands-on tests based on task samples drawn for each MOS. The MOS are shown, along with the numbers tested for each measure, in Table 1. The MOS are grouped into families following the classification by McLaughlin, Rossmeisl, Wise, Brandt, and Wang (1984).

Table 1

MOS, By Family, With N for Concurrent Validation

Family	MOS	SL1 Title	Written N	Hands-On N
Combat	11B	Infantryman	678	682
	13B	Cannon Crewman	639	619
	19E	Armor Crewman	459	474
Operations	31C	Single Channel Radio Operator	326	341
	63B	Light Wheel Vehicle Mechanic	596	569
	64C	Motor Transport Operator	668	640
Clerical	71L	Administrative Specialist	501	494
Skilled	91A	Medical Specialist	483	496
Technical	95B	Military Police	665	665

This paper focuses on the similarity of results for written and hands-on tests of tasks that were included for more than one MOS. The first section of the paper describes the communality among the MOS in the task domains and task samples. The second section considers the comparability of test results on the overlapping tasks.

Task Communality In Domains and Samples

One of the major decisions in the job analysis for Project A was whether criterion measures ought to maximize differences among jobs or seek to represent the range of task performance likely over a soldier's first enlistment. The decision was to represent the range of performance. This decision inevitably meant that there would be some communality among the criterion measures. The reason for the communality among measures is that there is considerable communality among the domains.

Army doctrine is that all skill level one soldiers are responsible for being able to perform the tasks in their MOS skill level one Soldier's Manual (SM) and the tasks in the Soldier's Manual of Common Tasks (SMCT). The task lists from these sources were automatically in the domain. Additional tasks were identified from analyses of Army Occupational Surveys. Most of the additional tasks were included in a higher skill level SM.

The amount of communality in the initial task domains is summarized in Table 2. Much of the communality results from including the common task list in each domain. Since the scope of the SMCT is subject to revision, the number of tasks varied by when the domains were defined. Domains for the first group of MOS--13B, 64C, 71L, and 95B--had 71 common tasks. The SMCT was revised before the second set of MOS were addressed, so MOS 11B, 19E, 31C, and 63B had 78 common tasks. MOS 91A (medical specialist) was also in the second group but claimed exemption from 12 weapons tasks, so their domain included 66 common tasks.

Table 2

Number of Tasks in Initial Task Domains

	11B	13B	19E	31C	63B	64C	71L	91A	95B	Overall
MOS	81	79	103	72	99	24	73	159	39	729
Shared	140	98	124	98	79	95	88	80	89	165
Total	221	177	227	170	178	119	161	239	128	894
% Shared	63	55	55	58	44	80	55	34	70	18

Since much of the overlap results from the common tasks, it is worthwhile to consider what the term "common task" means. These are tasks that are required of all soldiers regardless of job or location. The tasks have been identified and centrally managed by the Training and Doctrine Command on the premise that every soldier might be exposed to hostile action. The tasks themselves are not necessarily combat tasks. Some, such as first aid tasks, are also performed in non-lethal environments. Still the possibility that all soldiers might have to operate under combat conditions makes survival tasks as well as fighting tasks doctrinally required for all soldiers. Not all the communality among the MOS can be accounted for by a strict definition of common tasks. Some tasks are shared between MOS because soldiers operate similar vehicles (such as 1/4 ton truck), communicate under similar radio protocols, or are responsible for the same noncommon weapons (such as .50 caliber machinegun).

Overall there were 894 different tasks in the nine domains. Of these, 165 (18%) were shared by two or more MOS. The level of communality was naturally higher within each MOS domain--the average percent shared was 57.

The task communality in the domains extended to the samples selected for testing. Task selection was based on criteria of importance, difficulty, and frequency to represent content areas within the domain. The content areas were defined by having between 15 and 30 MOS supervisors per MOS sort all tasks in the domain by similarity. The supervisors also judged the importance of each task to the MOS mission and estimated a performance distribution for each task. The specific set of about 30 tasks to test per MOS was selected by a panel of six to nine people including subject matter experts and test developers. Initial selection proceeded by content area. Each selector independently chose a target number of tasks with the target being in the same proportion to 30 as the content area was to the total domain. Since common tasks tended to dominate consistent content areas [such as Nuclear, Biological, Chemical (NBC)] a high proportion of common tasks were selected.

The communality among tests is shown in Table 3. In this table a task is considered to be shared only if it was tested in the same mode for another MOS. The overall proportion of shared tasks (16% compared to 18%) is very similar to the proportion of shared tasks in the domains. However the average percent shared per MOS is less than in the domains (41 compared to 57).

Table 3

Number of Tasks Selected for Hands-On and Written Tests

Test Mode		11B	13B	19E	31C	63B	64C	71L	91A	95B	Overall
Hands-On	MOS	9	11	13	10	11	7	11	12	8	92
	Shared	5	6	2	5	4	9	4	3	8	12
Written	MOS	10	19	17	18	20	13	14	19	13	143
	Shared	18	11	12	12	10	15	11	11	18	32
Total	Tests	42	47	44	45	45	44	40	45	47	279
	MOS	19	30	30	28	31	20	25	31	21	235
	Shared	23	17	14	17	14	24	15	14	26	44
	%Shared	55%	36%	32%	38%	31%	55%	38%	31%	55%	16%

Another way of examining the degree of communality among the MOS is by means of the display in Table 4, which shows the percent of tasks from the domain of each MOS that are also found in the domain of each other MOS. Table 5 presents, similarly, the overlap among the tasks selected for each MOS. The correlation between the two levels of overlap--domain and selected task--is .51 across the 36 MOS pairs.

Table 4

Degree of Task Overlap Among MOS in Task Domains

	11B	13B	19E	31C	63B	64C	71L	91A	95B
11B	221	45.3	47.3	47.4	36.5	53.7	42.9	30.5	48.7
13B		177	38.7	42.7	42.3	56.2	45.1	30.5	43.1
19E			227	44.2	35.1	49.3	41.9	29.6	39.1
31C				170	39.7	55.0	46.0	34.2	43.8
63B					178	51.9	42.6	30.9	38.9
64C						119	59.2	41.5	56.8
71L							161	34.3	44.2
91A								239	33.0
95B									128

Note: Entries on the diagonal are the number of tasks in the MOS domain. Entries off the diagonal are indexes of overlap. Each index is the average percent of the two domains that overlap.

Table 5

Degree of Overlap Among MOS in Tasks Selected For Testing

	11B	13B	19E	31C	63B	64C	71L	91A	95B
11B	30	20.0	23.3	16.7	13.3	26.7	20.7	16.7	32.8
13B		30	13.3	16.7	16.7	26.7	13.8	20.0	16.4
19E			30	16.7	10.0	16.7	17.3	16.7	16.4
31C				30	16.7	26.7	27.6	30.0	26.2
63B					30	30.0	24.2	13.3	16.4
64C						30	24.2	20.0	23.0
71L							28	17.3	23.8
91A								30	23.0
95B									31

Note: Entries on the diagonal are the number of tasks selected for testing for each MOS. Entries off the diagonal are indexes of overlap. Each index is the average percent of tasks selected for the two MOS that overlap.

Comparability of Results on Shared Tests

The overlapping tasks are shown for each test method in Table 6 and Table 7. Since MOS are identified here only by job family, this table shows only how many MOS in each family had the test. (For these tables the clerical and skilled technical families have been combined.) All but three of the overlapping selected tasks are common tasks. The three exceptions are Set Headspace/Timing on .50, Perform Preventive Maintenance Checks and Services (PMCS), and Use Automated CEOI.

Table 6

Overlap Among Tasks Selected (Hands-On) Within Job Families

Category	Task	Combat	Operations	Clerical/ Skilled Tech
First Aid	Perform CPR	1	1	2
	Nerve Agent-Self	1	1	0
	Field/Pressure Dressing	2	3	3
Navigate	Grid Coordinate	1	2	3
	Azimuth-Mag. Compass	0	1	1
NBC	Put on M17 Mask	1	2	2
	Put on Prot. Clothing	1	1	1
Weapons	Op. Maint. M16	1	1	1
	Load/Reduce/Clear M60	1	1	1
	Set Headspace/Timing .50	2	0	0
	Load/Reduce/Clear M16	1	3	1
Vehicles	Perform PMCS	0	1	1

The task selection panels worked iteratively toward consensus that the tasks collectively represented the range of tasks required by skill level one soldiers in the MOS and individually were relevant to those soldiers' jobs. Senior officers and civilians at each proponent agency were asked to review the lists of tasks, again with an eye toward relevance. The final lists include their recommendations. Thus there is considerable defense for the proposition that all selected tasks are relevant. That is not to say, however, that every overlapping task is equally relevant regardless of MOS. For the combat family MOS, the potential of war has little effect on the relevance of tasks--their primary peacetime mission is to prepare for war. Similarly, first aid is at the core of daily operations of many medical specialists. For other MOS, however, the tasks do not fit as comfortably with daily operations, and maintaining proficiency on the tasks may interfere with their primary mission. Concerns about common task relevance have occasionally surfaced in the proponent agencies with the question "How common are the common tasks?" The overlap among tests gives a chance to look at the effect of MOS on test performance.

Table 7

Overlap Among Tasks Selected (Written) Within Job Families

Category	Task	Combat	Operations	Clerical/ Skilled Tech
First Aid	Perform CPR	2	2	2
	Nerve Agent-Self	3	1	1
	Field/Pressure Dressing	2	3	2
	Prevent Shock	2	0	1
	Nerve Agent-Buddy	0	2	0
Navigate	ID Terrain Features	2	0	0
	Navigate on Ground	1	0	1
	Estimate Range	1	0	1
	Grid Coordinates	1	3	3
	Azimuth-Mag Compass	0	1	2
NBC	Put on M17 Mask	2	2	2
	Put on Prot. Clothing	3	3	2
	Decontaminate Skin	1	2	2
	Maintain M17 Mask	1	1	1
Weapons	Op. Maint. M16	1	3	1
	Load/Reduce/Clear M60	1	1	1
	Set H/T .50	2	0	0
	Load/Reduce/Clear M16	1	3	3
	Op. Maint. .45	1	0	1
Field Tech	Call for Ind. Fire	1	0	1
	Collect/Report Info.	2	1	0
	Move Over Obstacles	1	0	1
	Camouflage Self	1	1	1
	Move Under Fire	1	0	1
	Install Claymore	2	0	0
	Camouflage Equipment	1	2	0
	Noise, Light, Litter Disc.	0	1	2
Communications	Use CEOI	1	0	1
Identify Targets	Identify Armor Vehicles	3	1	2
Customs & Laws	Know Rights as POW	1	1	1
	Challenge/Password	1	2	0
Vehicles	Perform PMCS	0	2	2

The differences among MOS were examined first by means of a one-way analysis of variance for each hands-on and knowledge test, with MOS membership as the independent variable. The ANOVA F values for the hands-on tests are shown in Table 8; the ANOVA F values for the written tests are shown in Table 9. As should be expected with the large number of cases, for almost every test, the overall test of differences among MOS is significant.

Table 8

Analysis of Variance Results on Hands-On Tests

Category	Task	Number of MOS	N	ANOVA F	P	Effect Size*
First Aid	Perform CPR	4	2413	387.56	.0001	.325
	Nerve Agent-Self	2	1252	4.34	.0374	.003
	Field/Pr. Dressing	8	4361	55.44	.0001	.081
Navigate	Grid Coordinates	6	3110	103.49	.0001	.143
	Azimuth-Mag. Compass	2	1234	21.56	.0001	.017
NBC	Put on M17 Mask	6	3662	10.24	.0001	.014
	Put on Prot. Clothing	3	1593	6.42	.0017	.008
Weapons	Op. Maint. M16	3	1816	199.21	.0001	.180
	Load, Clear M60	3	1987	744.03	.0001	.429
	Set H/T .50	2	1108	1.56	.2125	.001
	Load, Clear M16	5	2827	10.21	.0001	.014
Vehicles	Perform PMCS	2	1006	89.92	.0001	.082

*Calculated as (SS between)/(SS total)

The more pertinent question was whether any of the relationships was meaningful. To address this question a correlation ratio was calculated to determine how much of the variance could be attributed to MOS membership. As shown in the Effect Size columns, although most of the relationships are statistically significant, the proportion of the variance explained by MOS on most tasks is small, if not minuscule. The conclusion that we draw from these analyses is that, despite significant differences among MOS on the same tasks (largely due to the sample sizes), practical differences are rare.

Another way to determine how the MOS differ on the same tasks was to look at the distribution of mean differences. That was done for each test by comparing the means for each pair of scores. These results are summarized in Table 10 for hands-on tests and Table 11 for written tests.

Table 9

Analysis of Variance Results on Written Tests

		Number		ANOVA		Effect
Category	Task	of MOS	N	F	P	Size
First Aid	Perform CPR	6	1148	65.45	.0001	.087
	Nerve Agent-Self	5	2945	13.26	.0001	.018
	Field/Pr. Dressing	7	3875	19.63	.0001	.030
	Prevent Shock	3	1581	10.57	.0001	.013
	Nerve Agent-Buddy	2	1264	41.52	.0001	.032
Navigate	ID Terrain Feat.	2	1137	41.41	.0001	.035
	Navigate on Ground	2	1343	11.41	.0008	.008
	Estimate Range	2	1343	22.21	.0001	.016
	Grid Coordinates	7	3698	56.81	.0001	.085
	Azimuth-Mag. Compass	3	1762	12.66	.0001	.014
NBC	Put on M17 Mask	6	3747	5.97	.0001	.008
	Put on Prot. Clothing	8	4350	25.54	.0001	.040
	Decontaminate Skin	5	2781	5.57	.0002	.008
	Maintain M17 Mask	3	1492	.83	.4364	.001
Weapons	Op. Maint. M16	5	2769	12.99	.0001	.018
	Load, Clear M16	7	3878	7.27	.0001	.011
	Set H/T .50	2	1134	5.08	.0243	.004
	Load, Clear M60	3	2011	293.42	.0001	.226
	Op. Maint. .45	2	1123	.56	.4549	.000
Field Tech.	Call for Ind. Fire	2	1342	19.56	.0001	.014
	Collect/Report Info.	3	1805	66.94	.0001	.069
	Move Over Obstacles	2	1161	11.70	.0006	.010
	Camouflage Self	3	1844	58.91	.0001	.060
	Move Under Fire	2	1343	.54	.4636	.000
	Install Claymore	2	1137	112.22	.0001	.090
	Camouflage Equipment	3	1903	38.87	.0001	.039
	Noise, Light Disc.	3	1310	26.99	.0001	.040
Communications	Use CEOI	2	1124	.04	.8392	.000
ID Tgts.	ID Armor Vehicles	6	3250	164.55	.0001	.202
Customs & Laws	Know Rights as POW	3	1286	22.69	.0001	.034
	Challenge/Password	3	1903	7.83	.0004	.008
Vehicles	Perform PMCS	4	2142	79.44	.0001	.100

As an illustration of how to read the table, consider Load, Clear M60 in Table 10. Since three MOS were tested, three comparisons are possible. Two of the MOS did substantially better on the test than the third MOS. The difference between the high MOS is less than 1; the difference between each high MOS and the low MOS is in both cases more than 10. Such large differences are infrequent. Most of the hands-on comparisons show less than a five-point difference (59 of 87 comparisons). Most of the written comparisons show less than a 2.50 difference (125 of 213).

The conclusion from the examination of mean differences is similar to the conclusion from the mean effects--there is little practical difference between MOS performance on most tests. These results suggest that common tasks are, for most practical purposes, common.

Table 10

Distribution of Mean Differences Among MOS on Hands-On Tests

				1.0-	2.5-	5.0-	7.5-	Over
Category	Task	#MOS	<1	2.49	4.99	7.49	9.99	10
First Aid	Perform CPR	4	1		1		2	2
	Nerve Agent-Self	2		1				
	Field/Pr. Dressing	8	4	6	8	8	2	
Navigate	Grid Coordinates	6	2	3	2	4	4	
	Azimuth-Mag. Compass	2			1			
NBC	Put on M17 Mask	6	6	6	3			
	Put on Prot. Clothing	3	1	2				
Weapons	Operator Mnt. M16	3				2		1
	Load, Clear M16	5	4	4	2			
	Set H/T Cal .50	2	1					
	Load, Clear M60	3	1					2
Vehicles	Perform PMCS	2				1		
TOTAL			20	22	17	15	8	5

While most tests operate the same, there are cases where differences are evident. There was interest in looking at some tests to see the source of these differences. An Effect Size of .200 was selected as the criterion for that examination. The criterion is arbitrary: it was selected as a means of identifying extreme cases.

Table 11

Distribution of Mean Differences Among MOS on Written Tests

Category	Task	#MOS	<1	1.0- 2.49	2.5- 4.99	5.0- 7.49	7.5- 9.99	Over 10
First Aid	Perform CPR	6		4	6	3	1	1
	Nerve Agent-Self	5	2	5	3			
	Field/Pr. Dressing	7	4	9	7	1		
	Prevent Shock	3		2	1			
	Nerve Agent-Buddy	2			1			
Navigate	ID Terrain Feat.	2			1			
	Navigate on Ground	2		1				
	Estimate Range	2			1			
	Grid Coordinates	7	3	5	7	3	3	
	Azimuth-Mag. Compass	3	1	1	1			
NBC	Put on M17 Mask	6	7	7	1			
	Put on Prot. Clothing	8	6	8	10	4		
	Decontaminate Skin	5	6	4				
	Maintain M17 Mask	3	3					
Weapons	Operator Mnt. M16	5	3	4	3			
	Load, Clear M16	7	9	10	2			
	Set H/T Cal .50	2		1				
	Load, Clear M60	3		1			1	1
	Op./Mnt. Cal .45	2	1					
Field Tech.	Call for Ind. Fire	2			1			
	Collect/Report Info.	3		1	1	1		
	Move Over Obstacles	2		1				
	Camouflage Self	3		1	1	1		
	Move Under Fire	2	1					
	Install Claymore	2				1		
	Camouflage Equipment	3		1	2			
	Noise, Light Disc.	3	1		2			
Communications	Use Automated CEOI	2	1					
ID Targets	ID Armor Vehicles	6	4	2	2	2	1	4
Customs & Laws	Know Rights as POW	3		1	2			
	Challenge/Password	3	1	2				
Vehicles	Perform PMCS	4	1		4	1		
TOTAL			54	71	59	17	6	6

Two of the hands-on tests met the criterion. The MOS means and standard deviations for these tasks are shown in Table 12. The results have been standardized across all soldiers to a grand mean of 50 and standard deviation of ten. Each entry in the table shows the results for each MOS. Thus, two Clerical/Skilled Technical MOS were tested on CPR. In both cases the effect results from high performance in MOS where the task relates to the primary mission: with CPR medical specialists performed best; infantrymen and military police did best on the M60 machinegun.

Table 12

Means and SD for Shared Hands-On Tests With Effect Sizes >.200

Category	Task		Combat	Operations	Clerical/ Skilled Tech
First Aid	Perform CPR	Mean:	44.98	44.36	a. 57.98
		SD:	9.30	8.58	6.25
		Mean:			b. 54.10
		SD:			8.09
Weapons	Load, Clear M60	Mean:	54.80	40.51	54.21
		SD:	6.16	10.00	5.98

The written tests also had two tasks that met the criterion even though more tasks were covered in the written mode. The results for these tests are shown in Table 13. In both cases the results are consistent with intuition. As with the hands-on test, infantrymen and military police did substantially better than the soldiers in the operations MOS--though the effect was less for the written tests. It is also encouraging that MOS with an anti-armor mission (armor crewman and infantryman) are better at identifying enemy and friendly vehicles than other MOS.

Conclusions

There is considerable task communality among the criterion measures for the Project A MOS. That communality results largely from the common tasks that are required of all soldiers and is comparable to the communality in the task domains.

Most task tests yield similar results regardless of the MOS being tested. Where there are differences, they are more likely to occur in hands-on than in written measures. Extreme differences can be explained in terms of intuitive relevance to the primary mission of one or more MOS.

Table 13

Means and SD for Shared Written Tests With Effect Size >.200

Category	Task		Combat	Operations	Clerical/ Skilled Tech
Weapons	Load, Clear M60	Mean:	54.29	43.36	52.30
		SD:	9.45	8.34	8.56
ID Targets	ID Armor Vehicle	Mean:	a. 59.84	46.67	a. 47.31
		SD:	7.39	8.54	9.29
		Mean:	b. 52.23		b. 46.33
		SD:	9.26		9.13
		Mean:	c. 47.84		
		SD:	9.28		

REFERENCES

McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). Validation of current and alternative ASVAB area composites, based on training and SQT information on FY1981 and FY1982 enlisted accessions (ARI Technical Report 651). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

THE RELATION OF LEADERSHIP AND INDIVIDUAL DIFFERENCES TO JOB PERFORMANCE

Leaetta M. Hough
Personnel Decisions Research Institute

Ilene F. Gast and Leonard A. White
U.S. Army Research Institute

Rodney McCloy
Personnel Decisions Research Institute

Presented on symposium,
"Individual Differences, Work Environment, and Performance in Military Jobs"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

The Relation of Leadership and Individual Differences to Job Performance

The Army is involved in a selection and classification project (Project A) to develop new and validate current and new predictors of job performance during first tour of duty. The project involves gathering archival predictor and criterion data and developing and administering comprehensive sets of predictor and performance measures. The entire set of predictors, both archival and new, includes measures of cognitive ability, physical condition, temperament, interest, desired work environment, and perceived leadership constructs. The entire set of performance measures includes hands-on (work samples), knowledge tests, "administrative" indices (AWOL, discharge), and supervisory and peer ratings of specific job/task performance and Army-wide performance (effort, commitment). The project is an effort to study (1) the influence of individual difference characteristics measured prior to enlistment on job performance, (2) the influence of environment and Army work/life experiences on job performance, and (3) the nature of those relationships.

The focus of this paper is on perceived leadership and its relationship to individual difference variables and job performance. In an effort to understand perceived leadership constructs, we will examine the zero order correlations and multiple correlations between perceived leadership and individual difference variables, including cognitive ability, physical condition, temperament, interest, and desired work environment constructs. We will examine the size and the nature of the relationships between perceived leadership constructs and various aspects of job performance. We will also present a tentative model of the nature of relationships between perceived leadership constructs, general intelligence, temperament constructs, and various types of performance criteria.

METHOD

Subjects

Subjects were 5161 first-term soldiers in 9 military occupational specialties (MOS): 686 infantrymen (11B), 648 cannon crewmembers (13B), 492 tank crewmembers (19E), 353 radio teletype operators (31C), 624 light wheel mechanics (63B), 673 motor transport operators (64C), 504 administrative specialists (71L), 491 medical specialists (91A), and 690 military police (95B). Males comprised 88 percent of the sample and females 12 percent; 24 percent were black, 3 percent were Hispanic, and 68 percent were white; 4 percent indicated "other," and 1 percent indicated no racial origin. On the average, subjects had been in the Army for 20 months. Soldiers report of work experience in their unit ranged from 0 months to 48 months (median = 12).

Measures/Instruments

Perceived leadership. Critical incident workshops were conducted with NCOs in 5 Army MOSs: 11B, 19E, 31C, 64C, and 91A. These NCOs generated a total of 474 useable examples of leader behavior that influence performance of first-term soldiers. Items were written to capture categories of behavior represented in the incidents.

The questionnaire was field-tested in a sample of 696 first-term soldiers (White, ast, & Rumsey, 1985) and revised prior to administration in the present sample. Principal factor analysis (with promax rotation) of the items resulted in four factors that were descriptive of leader behavior. The four factors are Support/Inspiration, Structuring Work, Fairness/Discipline, and Participation. Scale scores were formed by summing the responses to the items loading highest on the factor. In addition, we formed an "overall quality of perceived leadership," which was the sum of the four scale scores. Table 1 shows the means, standard deviations, reliabilities, and the number of items for each of the four scales.

A fifth factor emerged that consisted of items describing work conditions over which supervisors in the Army have little control. These items were dropped for the present analyses.

Cognitive ability. The Armed Services Vocational Aptitude Battery (ASVAB), administered to all soldiers prior to entering military service, consists of 10 subtests. A composite measure of four of these subtests, known as the Armed Forces Qualification Test (AFQT), was used as the measure of general intelligence. In addition, as part of Project A, Personnel Decisions Research Institute (PDRI) developed six paper-and-pencil cognitive tests, which, when factor analyzed (principal factor analysis with varimax rotation), yielded the following three factors: Spatial Reasoning-Power, Spatial Reasoning-Speed, and Spatial Orientation (Toquam, Peterson, Rosse, Ashworth, Hanson, & Hallam, 1986). We used the AFQT and the three spatial ability factor scores as our measures of cognitive ability.

Temperament. During Project A, we developed 10 temperament scales and 4 response validity scales to tap important elements of temperament

Table 1

Descriptive Statistics of
Perceived Leadership Scales

Perceived Leadership Scale	No. of Items	Mean	SD	Reliability
Support/Inspiration	9	24.8	7.16	.87
Structuring Work	9	30.0	5.52	.79
Fairness	5	16.9	4.28	.75
Participation	4	12.9	3.17	.70
Overall Quality	27	84.8	16.08	.90

NOTE: Coefficient Alpha has been used as the reliability estimate.

Sample size is 5041.

Scale	Sample Items
Support/Inspiration	*Your supervisor understands your problems & needs.* Your supervisor wants to make you give your best effort.
Structuring Work	Before you start a task, you are told what has to be done & when it needs to be finished. Your supervisor follows up to make sure that assignments are completed.
Fairness	Your supervisor punishes you too severely. Your supervisor disciplines people without giving them a clear reason or explanation.
Participation	You are permitted to use your own judgment in solving problems. If you knew of a better way to do a task, you would feel free to share your ideas with your supervisor.

constructs that had demonstrated criterion-related validity in previous studies (Hough, 1984). An inventory entitled Assessment of Background and Life Experiences (ABLE) was prepared and pre-tested with a total of 470 soldiers at 3 separate forts. These data were used to revise the items and scales. Factor analysis (principal factor with varimax rotation) of approximately 8300 soldiers' responses to the ABLE indicated that the 10 ABLE temperament scales could be summarized with 3 factors--Achievement Orientation (Ascendancy), Dependability, and Emotional Stability (Hough & Ashworth, 1986). Factor scores on these three factors were used as the individual difference measures of important temperament constructs.

Physical ability. Also included in the ABLE inventory was a set of biodata type items asking about the respondents' physical condition prior to joining the Army. The score on this scale, "physical condition," was used in the present study.

Interest. During Project A, we also developed an interest inventory, entitled Army Vocational Interest Career Examination (AVOICE) that consisted of 23 scales. The AVOICE was pre-tested, along with ABLE, with 470 soldiers at 3 forts. These data were used to revise the items and scales. Factor analysis (principal component with varimax rotation) of approximately 7500 soldiers' responses to the AVOICE indicated that the 23 interest scales could be summarized with 6 factors--Skilled Technical, Structural/Machine Trades, Combat-Related, Food Service, Audiovisual Arts, and Protective Services (Hough & Ashworth, 1986). Factor scores on these six factors were used as the individual difference measures of important interest constructs.

Desired work environment. During Project A, we also developed an inventory entitled, Job Orientation Blank (JOB) to measure a person's desired work environment. It consisted of six scales. It, along with the ABLE and AVOICE, was pre-tested with 470 soldiers at 3 forts. These data were used to revise the items and scales. Factor analysis (principal component with varimax rotation) of approximately 7200 soldiers' responses to the JOB indicated that the 6 JOB scales could be summarized with 4 factors--Organizational Support, Coworker Support, Autonomy, and Service Others (Hough & Ashworth, 1986). Factor scores on these four factors were used as the individual difference measures of desired work environments.

Criterion Measures

Hands-on proficiency tests. For each of the jobs, 15 critical tasks were identified to represent the MOS-specific task domain. Multi-step task proficiency tests were prepared for each task. Each step of a task was scored pass or fail. A score for each task was computed by calculating the proportion of steps passed; these task scores were averaged to yield an overall hands-on test score.

Job knowledge tests. Important knowledge areas were identified through job analysis for each of the jobs. Subject matter specialists assisted Project A personnel in developing items to tap these knowledges. The overall knowledge test score was computed by taking the percentage of correct items.

Task/job performance (MOS-specific) rating scales. Behavior summary rating scales were developed for each job or MOS (see Pulakos & Borman, 1986). These instruments contained from 6 to 12 MOS-specific rating dimensions, each of which contained a 7-point scale ranging from 1 (low) to 7 (high). An additional scale required supervisors to make an overall assessment of task performance. Scores on this overall effectiveness scale were averaged across supervisors who provided ratings. This average formed an overall measure of task performance.

Army-wide performance rating scales. Ten behavior summary rating scales, each consisting of a seven-point scale ranging from one (low) to seven (high), were also developed to assess first-term soldier effectiveness in the Army (see Pulakos & Borman, 1986). These scales went beyond task performance to include aspects of socialization and commitment to the organization. An additional scale required supervisors to make an assessment of overall effectiveness. Scores on this overall effectiveness scale were averaged across supervisors who provided ratings. This average formed an overall measure of Army-wide effectiveness. In addition, factor analysis (principal component with varimax rotation) of the 10 Army-wide supervisor performance rating scales yielded 3 factors--Technical Skill and Effort, Integrity and Control, and Appearance. Factor scores on these three factors were also used as indicators of Army-wide performance.

RESULTS

Correlates of "Perceived Leadership" Scales

We correlated the four perceived leadership scales and overall quality of leadership scale with the 18 individual difference variables. As can be seen in Table 2, none of the cognitive ability variables, interest variables, desired work environment variables, or physical ability variables correlates in any important way with the perceived leadership measures.

Table 2

Correlations Between Perceived Leadership Scales and Individual Difference Variables

	Perceived Leadership				
Individual Difference Variable	Support	Structure	Fairness	Partici-pation	Overall Quality
Cognitive Ability Variables:					
General intelligence	-.02	.00	.05	.04	.02
Spatial reasoning-power	-.04	.02	.04	.05	.02
Spatial reasoning-speed	-.08	-.06	-.06	-.03	-.07
Spatial orientation	-.03	.00	.01	.00	.00
Temperament Variables:					
Achievement orientation	.07	.05	.01	$\boxed{.21}$	.11
Dependability	$\boxed{.21}$	$\boxed{.16}$	$\boxed{.21}$	$\boxed{.16}$	$\boxed{.23}$
Emotional stability	.11	.08	.13	.12	.13
Interest Variables:					
Skilled technical	.11	.07	.08	.11	.12
Structural/machine trades	-.04	-.01	-.10	-.01	-.04
Combat-related	.03	.06	-.03	.05	.03
Food services	.11	.04	.05	.03	.07
Audiovisual arts	-.02	.01	-.04	.01	-.01
Protective services	.00	.00	.01	-.03	-.01
Desired Work Environment Variables:					
Autonomy	-.05	-.03	-.08	.02	-.05
Organizational support	.05	.12	.11	.12	.12
Coworker support	.02	.12	.07	.08	.09
Serve others	.09	.09	.08	.11	.12
Physical Ability Variables:					
Physical condition	.00	.01	-.05	.02	-.01

NOTE: Sample sizes range from 3945-4976.
A box indicates correlations equal to or greater than .15.
Correlations equal to or greater than .04 are significant at $p\leq.01$.

Only the temperament variables correlate with the perceived leadership scales, and, of those, "dependability" is the most highly related to perceived leadership, ranging from .16 for "structuring work" to .23 for "overall perceived quality of leadership." The other temperament variable that correlates in an important way with perceived leadership is "achievement orientation;" its correlations range from .01 up to .21 with perceived "participation."

We next examined the extent to which scores on perceived leadership scales could be explained by combinations of other individual difference measures. Table 3 shows the cross-validated estimates of R^2 for each perceived leadership scale and combinations of individual difference measures. It shows the amount of variance in each leadership scale that can be explained by combinations of (a) temperament measures; (b) temperament and interest measures; (c) temperament, interest, and general intelligence measures; and (d) temperament, interest cognitive ability, physical condition, and desired work environment measures.

As can be seen, the perceived leadership scale that is best predicted with other characteristics of the subordinate is "participation," and it is the temperament of the subordinate, not his or her interests, cognitive abilities or other personal characteristics, that is predictive of the extent to which a subordinate describes his or her leader as "participative." The multiple correlation between temperament measures and "participation" is .33, the R^2 is .11. When all the individual differences measures are used to predict perceived

Table 3

Predicting Scores on Leadership Scales Using Other Individual Differences Variables

	Leadership Scale				
Predictor	Support	Structure	Fairness	Participation	Overall
Temperament measures	.07	.04	.08	.11	.08
Temperament & interest measures	.09	.05	.09	.12	.08
Temperament, interest, & general intelligence measures	.09	.05	.09	.12	.08
Temperament, interest, cognitive, physical, & desired work environment measures	.12	.08	.13	.14	.12

NOTE: Estimates of cross-validated R^2 are given.

Sample size is 2296.

"participation," the multiple correlation goes up only four points to .37 with R^2 equal to .14.

These data suggest that the subordinate's temperament is related to perceived leadership, but that the subordinate's interests, cognitive abilities, or other characteristics are not. The relationship between temperament and perceived leadership may be spurious in that the relationship might be due to a test-taking response set on the part of subordinates. The relationship, however, may not be spurious; it may, in fact, reflect a relationship between a subordinate's temperament and his or her relationship with the supervisor.

The next correlations we examined were the correlations between predictor composites and criterion composites. As can be seen in Table 4, perceived leadership scales do correlate in an important way with supervisory ratings of subordinate task performance and Army-wide performance, but they have little relationship with hands-on tests or job knowledge tests.

Similarly, the temperament variables correlate in an important way with supervisory ratings of task performance and Army-wide performance but not with hands-on tests or job knowledge tests.

Nature of the Relationship

We next examined the nature of the relationships between perceived leadership, various characteristics of the subordinate, and job performance. Some researchers (Barnes, Potter, & Fiedler, 1983) have suggested that leadership moderates the relationship between general ability and job performance. Other researchers (Schmidt & Hunter, 1977) have argued that the relationship between general ability and performance is stable across time and situations for similar jobs. We investigated

Table 4

Correlations Between Predictor Variables and Various Types of Criterion Measures

Predictor Composite	Criterion Composite			
	TESTS		SUPERVISORY RATINGS	
	Hands-On Tests	Job Knowledge Tests	Task Performance	Army-Wide Performance
Cognitive Ability Variables:				
General intelligence	.10	$\boxed{.42}$	.10	.09
Spatial reasoning-power	$\boxed{.17}$	$\boxed{.40}$	.08	.08
Spatial reasoning-speed	$\boxed{.14}$	.12	.06	.03
Spatial orientation	$\boxed{.17}$	$\boxed{.34}$	.06	.06
Temperament Variables:				
Achievement orientation	.04	.03	$\boxed{.15}$	$\boxed{.15}$
Dependability	-.05	.07	$\boxed{.12}$	$\boxed{.20}$
Emotional stability	.01	.11	.08	.08
Interest Variables:				
Skilled technical	-.06	.02	.03	.06
Structural/machine trades	$\boxed{.24}$	.04	-.01	-.05
Combat-related	$\boxed{.23}$	$\boxed{.23}$	.11	.08
Audiovisual arts	-.09	.00	-.02	-.01
Food services	-.13	-.11	-.05	-.03
Protective services	-.08	-.03	.02	.00
Desired Work Environment Variables:				
Coworker support	-.01	.09	.05	.06
Organizational support	.00	$\boxed{.13}$	.04	.06
Serve others	-.06	.05	.04	.07
Autonomy	.05	.06	.01	-.01
Physical Ability Variables:				
Physical condition	.02	-.04	.04	.05
Perceived Leadership Variables:				
Support	-.01	-.02	$\boxed{.14}$	$\boxed{.18}$
Structure	.01	.08	.06	.08
Fairness	.02	.09	$\boxed{.15}$	$\boxed{.23}$
Participation	.05	.07	$\boxed{.17}$	$\boxed{.21}$

NOTE: Sample sizes range from 3686 to 4996.
Boxes indicate the five highest correlates of each criterion.
Correlations equal to or greater than .04 are significant at $p \leq .01$.

these two competing hypotheses for various performance criteria. We divided the sample into thirds[1]--low, medium, and high--on the "overall quality" of perceived leadership scale and computed the correlations between general intelligence and various job performance criteria.

Table 5 shows the correlations. As can be seen, perceived quality of leadership does not moderate the relationship between general intelligence and performance criteria such as hands-on tests, knowledge tests or supervisory ratings of job performance.

We also examined whether perceived quality of leadership moderated the relationship between the three temperament variables and various types of performance criteria. Table 6 shows the results for "emotional stability," and Table 7 shows the results for "achievement orientation." None of the differences between correlations is significant. Table 8 shows the results for "dependability." Two of the differences are statistically significant at the .05 probability level. Perceived quality of leadership does appear to moderate the relationship between "dependability" and supervisory ratings of job performance. As can be seen, the relationship between "dependability" and supervisory ratings of job performance is highest when quality of perceived leadership is low. Stated in reverse, it appears that a soldier's personal trait of "dependability" is less important to ratings of their effectiveness when they are in units with better leaders. It is possible that less effective leaders emphasize "towing the line" in their ratings, and,

[1] We used this method rather than moderated regression analyses because Rorer (1971) found that in a data set generated to have moderator variables, the moderated regression analysis did not identify the moderator variables, whereas the split-group analysis did.

Table 5

Correlations Between General Intelligence[a] and Job Performance Within Different Levels of Perceived Quality of Leadership

	Overall Quality of Perceived Leadership	
Criterion Composite	Low Quality N = 1108	High Quality N = 1108
Hands-On Tests	.10	.11
Knowledge Test	.41	.42
Overall Task Performance	.05	.11
Army-Wide Performance:		
Technical Skill & Effort	.13	.17
Integrity & Control	.09	.10
Appearance	-.04	-.04

[a] General Intelligence as measured by the Armed Forces Qualification Test (AFQT, 1980).

NOTE: Differences in correlations of .08 or more are significant at $p \leq .05$.

Table 6

Correlations Between "Emotional Stability" and Job Performance Within Different Levels of Perceived Quality of Leadership

	Overall Quality of Perceived Leadership	
Criterion Composite	Low Quality N = 1108	High Quality N = 1108
Hands-On Tests	-.03	.04
Knowledge Test	.09	.14
Overall Task Performance	.11	.06
Army-Wide Performance:		
Technical Skill & Effort	.11	.04
Integrity & Control	.13	.09
Appearance	.08	.07

NOTE: Differences in correlations of .08 or more are significant at p≤.05.

Table 7

Correlations Between "Achievement Orientation" (Ascendancy) and Job Performance Within Different Levels of Perceived Quality of Leadership

	Overall Quality of Perceived Leadership	
Criterion Composite	Low Quality N = 1108	High Quality N = 1108
Hands-On Tests	.06	.07
Knowledge Test	.00	.06
Overall Task Performance	.13	.18
Army-Wide Performance:		
Technical Skill & Effort	.17	.24
Integrity & Control	-.02	.01
Appearance	.21	.15

NOTE: Differences in correlations of .08 or more are significant at $p \leq .05$.

Table 8

Correlations Between "Dependability" and
Job Performance Within Different Levels of
Perceived Quality of Leadership

	Overall Quality of Perceived Leadership	
Criterion Composite	Low Quality N = 1108	High Quality N = 1108
Hands-On Tests	-.04	-.04
Knowledge Test	.02	.07
Overall Task Performance	.11	.05
Army-Wide Performance:		
Technical Skill & Effort	.21	.10
Integrity & Control	.29	.24
Appearance	.21	.10

NOTE: Differences in correlations of .08 or more are significant at $p \leq .05$.

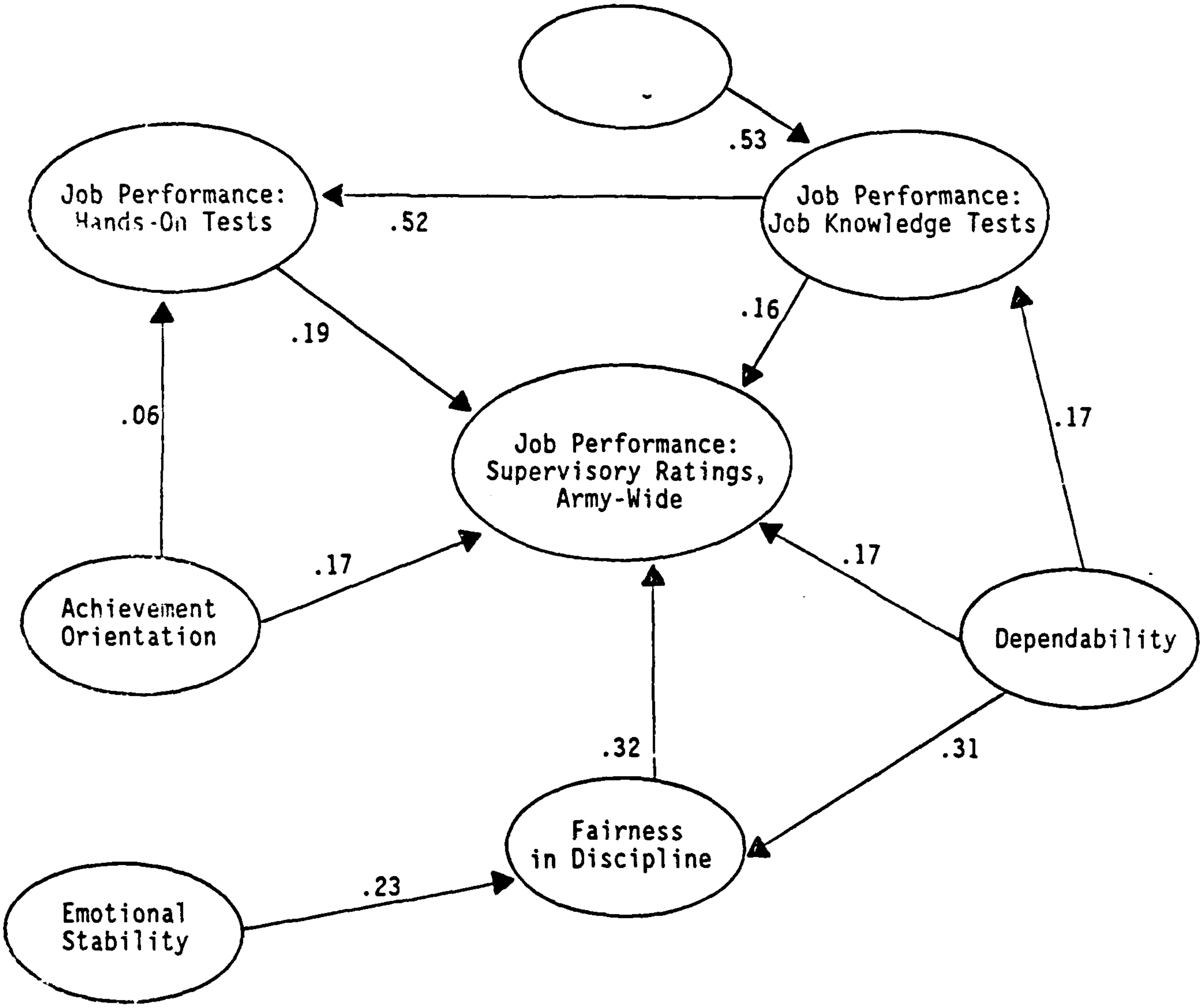
therefore, the "dependability" factor shows up as more important under those conditions. Recall, however, that "dependability" correlates positively, .23 with perceived quality of leadership; thus, this interpretation is not likely. These data do not, of course, indicate causal links; they do, however, indicate that the trait "dependability" is less strongly related to supervisory ratings of performance under conditions where leadership is seen as more effective.

We have also done some preliminary path analyses. Tables 9 and 10 show what are, at this point, tentative models of the relationships between two of the perceived leadership scales, subordinate individual differences, and various types of job performance criteria. The analyses were performed on half the data from three combat MOS (11B, 13B, and 19E); we have not yet had a chance to conduct path analyses on the other half of the data set. Nor have we had a chance to perform path analyses on the other two leadership scales. Nevertheless, we would like to present our preliminary models.

Table 9 shows the tentative model that includes perceived "fairness" of leader's discipline. It represents the third iteration and is, thus, a trimmed model. The goodness of fit, .987, indicates the model is a good fit with the data. The perceived "fairness" of leader's discipline has a significant path coefficient to supervisory ratings of Army wide performance. It is also the largest path coefficient that goes to the supervisory ratings. The subordinates' "dependability" and "emotional stability" have a significant path coefficient to perceived "fairness" of leader's discipline and seem to exert most of their influence on supervisory ratings via that avenue. The top portion of the model is similar to the model Hunter (1983) presented, although we

Table 9

Tentative Model of the Relationships Between Perceived "Fairness" in Leader's Discipline, Subordinate Individual Differences, and Various Types of Job Performance Criteria



.53
Job Performance: Hands-On Tests
.52
Job Performance: Job Knowledge Tests
.19
.16
.06
Job Performance: Supervisory Ratings, Army-Wide
.17
.17
.17
Achievement Orientation
Dependability
.32
.31
Fairness in Discipline
.23
Emotional Stability

X^2(15df) = 17.67; p = .28
GF = .995; .987 adjusted for d.f.
Coefficient of Determination = .436

NOTE: Sample consists of three combat MOSs; N-800 found somewhat lower path coefficients from both hands-on tests and job knowledge tests to supervisory ratings of job performance.

The next table, Table 10, shows the tentative model for perceived "participatory" leadership, subordinate individual differences, and various types of job performance criteria. As with the last model, this too is a trimmed model. The top half of the model is, of course, the same as the previous model. In this model, perceived "participatory" leadership and "dependability" have a significant and similar size path coefficient to supervisory ratings, hands-on tests, and, to a lesser degree, job knowledge tests and "emotional stability," also a significant path coefficient to supervisory ratings of job performance. "Dependability," "achievement orientation," and "emotional stability" all have significant path coefficients to both perceived "participatory" leadership and supervisory ratings, though "achievement orientation" appears to exert most of its influence on supervisory ratings via perceived "participatory" leadership.

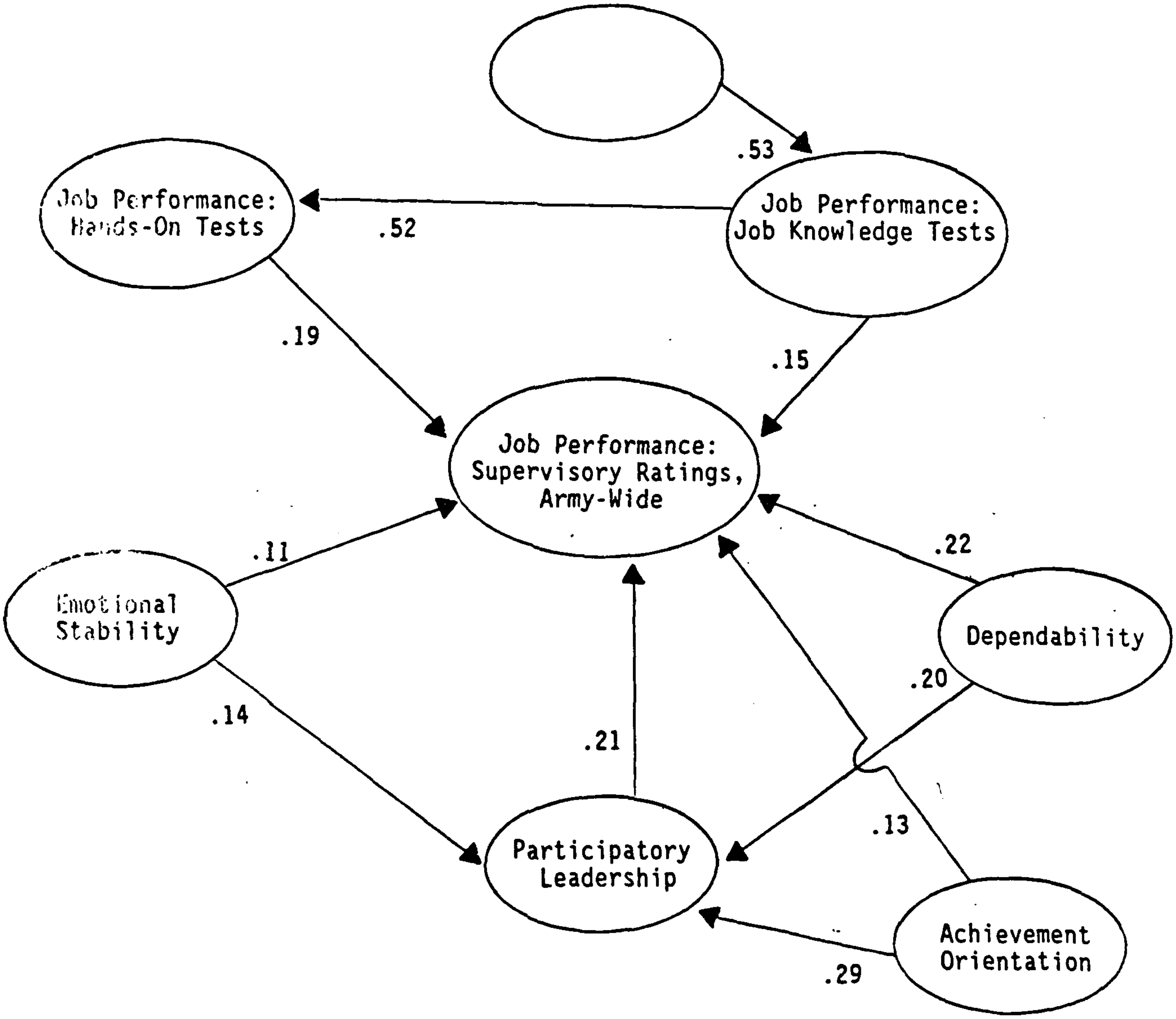
As mentioned earlier, these models are tentative; we intend to refine them and do other path analyses that include the two other perceived leadership scales and specific constructs of supervisory ratings.

SUMMARY AND DISCUSSION

The present research is a departure from the traditional approach to studying leadership. In the past, research has often emphasized the traits, characteristics, and behaviors of individuals in leadership positions. In the past 15 or so years, attention has been given to the process or interaction between the leader and other group members. Considerably less attention has been given to the enduring personal characteristics, except, perhaps, for general intelligence, of the

Table 10

Tentative Model of the Relationships Between Perceived Participatory Quality of Leadership, Subordinate Individual Differences, and Various Types of Job Performance Criteria


.53
Job Performance: Hands-On Tests
.52
Job Performance: Job Knowledge Tests
.19
.15
Job Performance: Supervisory Ratings, Army-Wide
.11
.22
Emotional Stability
Dependability
.20
.14
.21
.13
Participatory Leadership
Achievement Orientation
.29

X^2(14df) = 14.97; p = .380

GF = .995; .998 adjusted for d.f.

NOTE: Sample consists of three combat MOSs; N-800 subordinates. Our approach has focused on the subordinates--his or her perceptions of quality of leadership, his or her personal characteristics, and his or her job performance as indicated by various criteria.

Our data suggest:

(1) That perceived quality of leadership variables are uncorrelated with most individual difference variables such as cognitive abilities, physical condition, interests, and desired work environments;

(2) that perceived quality of leadership variables are correlated with the temperament variables, especially the "dependability" trait;

(3) that perceived quality of leadership moderates the relationship between the temperament trait "dependability" and supervisory ratings or job performance;

(4) that perceived quality of leadership does not moderate the relationship between cognitive ability and job performance; and

(5) that quality of leadership variables do correlate with supervisor ratings of job performance.

We have also developed two tentative models of the relationships among the various measures that suggest:

(6) that perceived quality of leadership has a direct influence on supervisory ratings of job performance; and

(7) that the subordinate's temperament appears to have a direct influence on supervisory ratings of job performance, but also

(8) that the subordinate's temperament appears to have an indirect influence on supervisory ratings of job performance that is expressed via perceived quality of leadership.

REFERENCES

Barnes, V., Potter, E. H., & Fiedler, F. E. (1983). Effect of interpersonal stress on prediction of academic performance. Journal of Applied Psychology, 68, 686-697.

Hough, L. M. (1984). Identification and development of temperament and interest constructs and inventories for predicting job performance of Army enlisted personnel. Minneapolis, MN: Personnel Decisions Research Institute.

Hough, L. M., & Ashworth, S. D. (1986). Project A concurrent validity data analyses: Temperament and interest predictors. Minneapolis: Personnel Decisions Research Institute.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. Performance measurement and theory (pp. 257-266). New Jersey: Lawrence Earlbaum Assoc.

Pulakos, E., & Borman, W. C. (1986). Project A: Developing the basic criterion scores for ratings. Minneapolis: Personnel Decisions Research Institute.

Rorer, L. G. (1971). A circuitous route to bootstrapping selection procedures. In Personality measurement in medical education, (ed.), H. B. Haley, A. G. D'Costa, & A. M. Schafer. Des Plaines, IL: Association of American Medical Colleges.

Schmidt, F. L., & Hunter, J. E. (1977). The development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Toquam, J., Peterson, N., Rosse, R., Ashworth, S., Hanson, M., & Hallam, G. (1986). Project A concurrent validity data analyses: Cognitive paper-and-pencil and computer-administered predictors. Minneapolis: Personnel Decisions Research Institute.

White, L. A., Gast, I. E., & Rumsey, M. G. (1985). Leader behavior and the performance of first-term soldiers. Paper presented at 92nd Annual American Psychological Association Convention, Los Angeles, CA.

STABILITY AND INSTABILITY OF INDIVIDUAL DIFFERENCES

Lloyd G. Humphreys

Project A Scientific Advisory Group

Presented on symposium,
"Promising Areas of Psychometric Research"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

STABILITY AND INSTABILITY OF INDIVIDUAL DIFFERENCES

Lloyd G. Humphreys

University of Illinois

In 1960 I published a paper entitled Investigations of the Simplex that was concerned with the changes in individual differences relative to a group of people (age group, class in school, sample used in research) with the passage of time. Time, of course, merely allows various kinds of processes to take place (growth, decay, learning, forgetting, fatigue, response to incentives). In one way or another I have been interested in correlation analysis of longitudinal data in the years since.

Guttman had coined the term simplex to describe the intercorrelations of binary items in a perfect Guttman scale and to the intercorrelations of tests having the same content and differing in complexity administered on a single occasion. In such cases correlations are highest between items adjacent in difficulty or between tests adjacent in degree of complexity. Correlations decrease in magnitude as items depart from each other in difficulty or tests depart from each other in complexity.

A perfect simplex as defined by Guttman assumed error-free measurement and population correlation matrices. Correlations in a sample among fallible measures that are similar to a perfect simplex are called quasi-simplex matrices. In a true simplex all partial correlations between two remote measures in which an intermediate measure is held constant are equal to zero. This implies that gains or change between adjacent measures are independent of the initial score. Most such partial correlations in a quasi-simplex are positive. By obtaining estimates of reliabilities it is possible to fit the simplex model to observed correlations but this does not rule out the possibility that several other models could be fit to the same data. All such models would reduce the size of the positive partial correlations. The existence of a quasi-simplex does, however, narrow the choice among models.

Examples of Quasi-Simplex Matrices

Finding a quasi-simplex matrix indicates unequivocally that the correlations cannot be described by a single common factor when the approach is through principal factor analysis, but nonetheless there may be a single common factor underlying the correlations. The single factor in the quasi-simplex matrix is confounded with an ongoing change process. This is seen most readily in the intercorrelations of dichotomous items composing a test. If the items measure only one factor, the simplex model will fit the observations. In this case no other model will do as well. The confounding process is the change in difficulty levels of the items. After allowing for measurement error the correlation between two items that are close in level of difficulty will approach 1.0, but the maximum possible product-moment correlation (phi coefficient) between items far removed in level of difficulty from each other is substantially less than 1.0.

Several tests involving numerical content but ranging in complexity from simple addition to arithmetic reasoning form a quasi-simplex. The tests do measure a common factor of content, but differences in complexity produce differences in the correlations of the component data with each other and with other measures. One can find a similar quasi-simplex among verbal tests. All components of the matrix measure verbal content, but the correlation between arithmetic reasoning and verbal reasoning will be higher than the correlations between simple numerical and complex verbal, or between simple verbal and complex numerical.

As Fleishman has amply demonstrated, the intercorrelations of trials, or blocks of trials, in motor skills learning form quasi-simple:: matrices. A measure of skill obtained early becomes more and more fallible as a predictor of later skills as the number of intermediate trials increases. People change their relative standing on the task as practice continues. People are learning so that there is clearly change within the person, but this learning also produces changes in correlations of various predictors of performance, administered before the first learning trial, with early and late trials. The factorial composition of the task changes as learning progresses.

No motor skills task on which learning occurs measures aptitude in the traditional meaning of that construct. A given standard score depends on the phase of learning during which the score is obtained, the knowledge and skills brought to the learning task, and various aspects of the learning situation. For example, differences in motivation to do well on the task are probably involved in the variance of individual differences late in learning. The learning task in the objective sense remains constant, but scores have differential meaning as a function of the confounds described.

The intercorrelations of scores on intelligence tests administered over several occasions also form quasi-simplex matrices. The amount of change in scores relative to the age group is greater in the preschool period than later, but change does not stop. With good data the obtained correlations can be fit quite satisfactorily by the simplex model, but other data indicate that the model cannot be precisely correct. That it is reasonably close to the real world, however, is substantiated by the small observed correlations between mental age at time 1 and the gain in mental age from time 1 to time 2. These near zero correlations have been in our literature for more than 45 years.

The stability of IQs from year to year is quite high. Estimates of the correlation between true scores over a single year are in the nineties from the late preschool period on, but with somewhat lower stabilities in the earlier years than in the later years of development. Prepubescent height is somewhat more stable than intelligence, but during the adolescent growth spurt the true score stabilities are more nearly equal.

It is clear that neither height nor IQ is a stable trait of the organism. It is possible for persons who are not thinking clearly to claim that intelligence is really stable even though tested intelligence is not, but one cannot retain a construct in science that can neither be measured nor inferred from measurements. I have seen claims such as the following. Children given special educational treatment showed an increase in test performance that later dropped back to the pretraining level, but the author concluded that their IQs did not change. Such conclusions are statements of faith, not of testable psychological theory.

Example of Needed Research

If a necessary characteristic of a psychological trait is being fixed at a stable level over long periods of time, there are probably no psychological traits. (A psychological construct requires more than faith.) Most traits of physique or of physiology are also only relatively stable over short time periods when shortness is evaluated against the life span. The human organism is dynamic. Instability occurs during normal development without experimental intervention. The research questions that are of immediate concern are the degrees of stability manifested by the various traits that psychologists are interested in. For example, is fluid intelligence more or less stable than crystallized intelligence? How does the degree of stability of these two broad factors relate to the stability of intellectual speed, a third broad factor.

The next question in priority relate to the causes of change in individual differences. For technical reasons I believe that it is necessary at this point in history to focus on environmental causes, but I do not rule out possible genetic causes. Genes are not influential once and for all at the time of conception. They can and do "fire" at different times during development. The instability of both height and general intelligence may well be in part genetically determined.

Knowing that the correlation between estimated true scores on a standard test of intelligence are correlated to the extent of perhaps .97 from one year to the next does not make the search for causes likely to produce positive outcomes. On the other hand, the stability of those scores between 8 and 18 (in terms of estimated true scores again) is probably close to $.97^{10}$ which provides a considerably larger segment of true score variance to reflect causes of change. Therefore, the search for causes of change in intelligence is necessarily a long term proposition.

In contrast, in some recent military data the reliability of simple reaction time at time 1 is very high, but in a period of 2 to 4 weeks the correlation between observed scores shrinks below .50. Simple reaction time at any given point in time cannot be a very important psychological characteristic in general, and especially not a useful one for prediction purposes, unless the causes for this degree of instability can be found and controlled. Speed of information processing is an attractive construct, but much research is needed.

Prediction of Academic Grades

A number of years ago I published correlations between a measure of the high school academic record and scores on the American College Test as predictors and the 8 successive semesters of college grade point average, with these criteria being computed independently for each semester. There is a marked decline in the predictive correlations from the freshman to the senior year. The intercorrelations of these grade averages do also form a quasi-simplex so that the prediction of senior grades from freshman grades is highly fallible as well. Since then this result has been replicated by others many times. I have also contributed additional information. The Verbal and Quantitative scores on the Graduate Record Examination administered in the senior year also have their highest correlations with freshman grades and lowest with senior grades. The predictive and postdictive validity gradients are almost identical. (Advanced tests have their highest correlations with sophomore and junior grades.) Selection tests administered in the senior year have their highest correlations with first year graduate and professional school grades. The intercorrelations of these grades, of course computed independently, also form quasi-simplex matrices.

One indication of what is happening is furnished by a comparison of samples of high and low academic promise. The low promise sample showed substantially less stability of academic grades from semester to semester and correspondingly lower correlations with predictors. There was also a good deal more attrition over the four undergraduate years in the low promise group. Presumably they did a good deal more shopping around for easy curricula and easy instructors, and the ease of doing this in any given institution is probably related to the degree of instability of grades. It is also probable that the low promise students were highly selected on noncognitive traits related to academic success. It is also known that there is true score change in individual differences on broad cognitive tests from 18 to 22, but the amount is much smaller than the change reported for academic grades.

Ability Grouping. If children are placed in heterogeneous groups in school, or even if grouping varies from subject to subject and from time to time, children meeting some standard of gifted at age 6 will be closer to the population mean at 18 than at 6. By the same reasoning students who are retarded, but without organic etiology, at age 6 will be closer to the mean at 18. In both cases the regression will be by a substantial amount (for a close approximation) start with a true correlation between 6 and 18 of .96 or .97 to the 12th power). What will happen if the gifted group is segregated from other children and provided with the most effective curriculum the schools can provide? Will the mean of the segregated group regress toward the population mean or will effective segregated education maintain the level of the group at something closer to the level at 6? The same possibilities apply to the retarded group if they are provided a retarded curriculum. Will placement in classes for the educable mentally retarded decrease the regression of the group mean to the population mean? Within both groups it is highly certain that quasi-simplex matrices will represent the intercorrelations over the years, but the research question is the location of the mean at 18 toward which individuals will regress. The mean may change as a result of the intervention.

There are generally accepted quality differentials between institutions of higher education, but it has been quite difficult to demonstrate differentials in achievement when the quality of the admitted freshmen is taken into account. Is there actually very little that the institution accomplishes academically? To the person who accepts the reality of change of individual differences over time, another perspective is possible. If students were assigned at random to colleges, means would differ only by chance and individual students would regress toward the population mean. Even this mean, however, might well be the mean of entering freshmen rather than the unselected population mean. When the means of graduates of selective colleges do not regress toward the all-freshman mean, this alone is evidence for a positive institutional effect. When, in addition, investigators are able to show positive effects by the usual standards, this is an increment to the effect of the quality of the institution.

Criterion Instability. Independently computed grade point averages in most educational institutions show marked change, and there may well be many different causes for the change. There is every reason to believe that the same phenomenon will be found for performance criteria in industry and the military, and again there may be quite diverse causes at work. If the same sample is studied on several occasions, it is highly probable that the tests used in selection will be more valid for early than for late performance. A decrease in the size of the validity coefficient, perhaps to a trivial level, is not in itself sufficient evidence that the nature of the job has changed. It may not be possible to find a predictor with higher correlations with late performance. This perspective on stability and instability in test and criterion performance suggests that the validity of a selection program can only be validated by comparing the mean performance of the selected group with that of an appropriate control. Selection creates ability groups.

Stability and Instability in Race Differences

Both black and white students produce simplex matrices of measures of cognitive ability over time. Over long enough time intervals a great deal of change is found and in about equal amounts. With the possible exception of the years since about 1970, however, the individual regression was about the racial mean. Differences between the group means remained approximately constant in most schools from the first to the 12th grade. (Jensen found that blacks did show progressively lower means in some segregated southern rural schools.) Do blacks and whites in integrated classrooms regress toward the joint mean or toward their respective racial means? Has segregation been largely responsible for the constancy of the racial difference in the first 12 grades?

I introduced a caveat in the preceding discussion. Since 1970 blacks have made gains in reading comprehension, a skill that is near the core of general intelligence. Black 9-year olds made the first big gain between the 1971 and 1975 cycles. Four years later the same birth cohort made a significant gain as 13-year olds and 8 years later repeated the gain at 17. The one missing piece of information that provides a basis for caution concerning school effects is the level of scores on an intelligence test at the time of school entrance. It is conceivable that black 9-year olds in 1975 had higher scores on intelligence test than their predecessors at the time they entered school. (An intelligence test measures the knowledge and skills acquired in the home and neighborhood during the preschool period.) Such data are critical, especially given the historic stability of the race difference, in trying to develop programs for further gains. Between the two world wars both blacks and whites gained without changing appreciably the mean difference, but starting in the midsixties, when the 1975 9-year old black children were born, conditions for blacks had changed independently.

INDIVIDUAL DIFFERENCES AND ENVIRONMENTAL DETERMINANTS OF ARMY PERFORMANCE CRITERIA

Darlene M. Olson
U.S. Army Research Institute

Walter C. Borman and Stephen J. Motowidlo
Personnel Decisions Research Institute

Presented on symposium,
"Individual Differences, Work Environment, and Performance in Military Jobs"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

AUTHOR NOTES

The research described in this paper was conducted as a component of Project A (Contract No. MDA 903-82-C-0531), the Army's comprehensive and innovative soldier selection and classification project. Prior to Project A, Army selection and classification tests were related to school success, but had not been linked explicitly to job performance. Further, previous research had examined only the effects on job performance of abilities, personality characteristics, and motivation. Other factors such as work environment constructs and organizational variables may enhance the prediction of soldier performance or more thoroughly describe the interrelationships among human attributes, enlistment standards, selection criteria, and job performance.

For over 50 years, research in the areas of personnel selection and classification, and organizational psychology has attempted to conceptualize, describe, predict and measure performance in diverse work environments. Job performance has been conceptualized as a product of personal attributes/characteristics, abilities, and skills which are measurable at the time an individual first enters the organization, of environmental/organizational variables which impact on the individual after job-entry and of the person's motivation to perform. Historically, job performance has been studied in terms of taxonomies of human cognitive abilities, values, vocational interests and personality dimensions (Dunnette, 1976; Campbell and Pritchard, 1976), with extensive validation research aimed at predicting job performance from individual difference measures. Although such prediction research has found significant validities, only a portion of the total variability in performance criteria has been explained by individual differences. Despite considerable research in performance measurement, until recently little emphasis has been placed on the multidimensionality of the performance domain. Further until the pioneering work of Schneider (1975; 1983), Peters and O'Connor and their colleagues (1980; 1984), and James and his colleagues (1974; 1978; 1982), a paucity of research addressed the development of environmental and organizational climate taxonomies or examined relationships between these variables and work-related outcomes.

Work Environment

One major class of variables that may influence work performance is the work setting or environment. The work environment serves as the context in which performance occurs (Magnusson, 1981). Specifically, situational and environmental factors have been defined as a set of conditions/circumstances that are likely to influence the behavior of at least some individuals and have a reasonably high probability of reoccurrence in essentially the same form. (Frederiksen, Jensen, & Beaton, 1977).

Although the work environment provides the context and opportunity (or lack thereof) for performance-based behavior, individuals are not passively shaped by environmental contingencies. Rather, individuals actively process environmental information, develop strong perceptions and attitudes toward existing and previously experienced events, and are goal-directed participants in an ongoing reciprocal person by situation interaction process (Bandura, 1978; Magnusson & Endler, 1977). To explain work performance more effectively it is necessary to identify and measure reliably the relative influences of individual differences and environmental factors.

Environmental Constraint Research. Laboratory research conducted by Peters and O'Connor and their associates has demonstrated that situational constraints are significantly correlated with ineffective task performance, job dissatisfaction and increased frustration (e.g., Peters, O'Connor, & Rudolf, 1980). Results from initial field studies conducted on both civilian managers (O'Connor, Peters, Pooyan, Weekley, Frank, & Erenkrantz, 1984) and Air Force enlisted personnel (Watson, O'Connor, Eulberg, & Peters, 1983) have revealed weak but sometimes significant correlations between overall environmental constraints and supervisory performance ratings.

In other research conducted on Army enlisted personnel, Olson and Borman (1986) examined the relationships between Army work environment dimensions and a comprehensive set of performance measures. Significant correlations (ranging from the low to mid .20's) between supervisory ratings of overall soldier effectiveness and NCO potential, and such environmental variables as Individual Support, Role Models, and the Organizational Reward System were observed. Also, a statistically significant ($p < .05$) correlation of .22 was found between objective hands-on test performance and the work environment dimension of Training. Further, a significant negative correlation ($r = -.27$) was noted between job knowledge test scores and the environmental variable related to Resources/Tools/Equipment.

More substantial associations between the work environment and performance were evident in a field study conducted by Steel and Mento (1986). Results from their research showed significant effects of high vs. low environmental constraints on supervisory appraisals ($r = -.36$), self-ratings ($r = -.31$), and a measure of objective performance ($r = -.12$).

Generally, the environmental constraint research has provided valuable information on the relationships between work environment variables and indices of job performance. However, the large number of low correlations and non-significant results reported for this type of research suggests that the magnitude of the correlation coefficients may be dependent on: 1) the level of facilitating and inhibiting conditions actually present in the work environment, 2) the manner in which situational/environmental variables are conceptualized, 3) the kinds of jobs investigated, 4) the types of performance measures examined, 5) the potential influence of a wide array of unmeasured individual difference factors, and 6) the psychometric properties of the actual measures employed in the research.

Therefore, in order to describe performance more effectively and begin to understand the multidimensionality of the performance domain, it is important to develop conceptual models. These models could be used to investigate the complex interrelationships among individual differences, environmental factors, and a comprehensive set of performance criteria, including ratings and more objective maximum performance. Although such an expanded research emphasis thrusts one into the center of the venerable 'trait-situation' controversy, it is perhaps the only viable way to understand the complex influences of person and environment factors on job performance. That is, to understand human behavior in work settings, we must attend to both the person and the situation. Whether stable personality traits, the stimulus and contextual characteristics of situations, or some interaction of these factors determine individual work behavior should be considered in conceptual model building. If work environments and individuals are interdependent, then reciprocal determinism may help account for some variability in performance. For example, individuals with certain temperaments and/or dispositions may cause the environment through maintaining different perceptions of and responses to that environment. Also, environmental situations encountered on the job can control person factors through the active manipulation of reinforcement contingencies.

Conceptualization of a Model of Soldier Perfomance. Guidance for the development of a model of the influences on soldier work performance can be obtained from several sources. First, in the theoretical arena, Borman (1983) has discussed the implications of the trait-situation controversy for performance measurement research. Borman contends that consistency in individuals' work performance is greater than consistency of behavior in general because: 1) workers typically encounter a substantially restricted sample of environmental situations on the job that are more repetitive than those found in everyday life, and 2) relatively stable cognitive abilities comprise performance in the majority of jobs. This line of reasoning suggests that perhaps abilities and dispositional tendencies or personality variables would play a greater role in determining work performance than situational factors. However, Borman suggests that some situationally-based variability in performance is very likely to occur on the job and it is important to measure that source of influence.

Second, in a meta-analysis of 14 research studies, which investigated the relationships among three variables: ability, job knowledge, and performance (measured with work sample tests and supervisory ratings), Hunter (1983) conducted a causal analysis which showed that factors other than job performance and job knowledge explain a large portion of the variation in performance ratings. Examination of Hunter's 1983 model reveals a high correlation between ability and job performance that was partially related to the direct impact of ability differences on performance but that was more the result of an indirect influence due to the high correlation between ability and job knowledge. Also, a moderately high correlation was observed between job performance and supervisory ratings, but this was somewhat determined by the extent to which supervisors are sensitive to differences in job knowledge.

This joint-influence model of Hunter shows that ratings are influenced more by the knowledge workers have about their jobs than by how well they can, under standardized conditions actually, perform their jobs. Since ratings are more indicative of typical performance measures, it seems logical and reasonable that ratings would be less influenced by work samples, which are maximal performance criteria. The challenge for future research is not only to replicate, but to expand the Hunter model to include previously uninvestigated variables related to ratee characteristics and contextual/situational factors operating in work environments.

In response to Guion's (1983) call to expand the kinds of variables investigated in the Hunter model, Schmidt, Hunter and Outerbridge (1986) examined the effects of level of job experience on job knowledge, performance capability (i.e., measured by job sample tests) and supervisory performance ratings. They found that job experience (*M* of 2 to 3 years) has a direct impact on job knowledge and a smaller direct influence on job sample test performance. Also, job experience was observed to have an indirect relationship with work sample performance through an effect on job knowledge, which was subsequently noted to have the strongest effect on the work sample measures. Generally, the pattern/magnitude of causal relationships related to ability were similar to job experience, and overall the findings of Hunter (1983) were supported.

Finally, two recent empirical studies conducted by Pulakos and Schmitt (1983) and Staw and Ross (1985) suggest that job attitudes and satisfaction are related more to the dispositional state of the individual as opposed to being situationally-based. Pulakos and Schmitt found that preemployment expectations addressing the extent to which a job will meet existence, relatedness and growth needs were positively correlated (_r_ = .11 - .28) with subsequent job satisfaction. Staw and Ross, in their analysis of a large national sample of data on males' job satisfaction, provided considerable support for the dispositional argument that job attitudes are consistent within individuals, and show stability both over time and across situations. Since these current research studies found dispositional effects for job satisfaction criteria, it is conceivable that there may well be dispositional sources of variance in job performance.

Although the previously discussed theoretical and empirical perspectives appear to support a more trait or dispositional approach to understanding criterion variance, a comprehensive model of job performance must also consider situational influences as well as person by situation interactions (Schneider, 1983).

The purposes of the present research are to: 1) continue examining the magnitude of direct relationships between Army work environment factors and measures of both typical (e.g., rating factors) and maximal (e.g., job knowledge and hands-on tests) performance, 2) develop an exploratory path-analytic model to test the interrelationships among individual differences (e.g., ability and temperament), environmental factors, and performance criteria, and 3) extend the work on performance models initiated by Hunter (1983), through using multiple measures of the typical performance domain (e.g., Army-wide rating factors: Technical Skill and Effort, Integrity and Control, and Appearance, and Overall Soldier Effectiveness ratings) not just supervisory ratings of performance.

METHOD

Sample

Subjects in this research were 5080 first-term Army enlisted personnel in 9 different Army jobs. These soldiers were sampled from an array of military occupational specialties (MOS): the 11B (Infantryman), 13B (Cannon Crewman), 19E (Armor Crewman), 31C (Radio Operator), 63B (Light Wheel Vehicle Mechanic), 64C (Motor Transport Operator), 71L, (Administrative Specialist), 91A (Medical Care Specialist), and 95B (Military Police) at multiple Army installations in the Continental United States (CONUS) and Europe. Table 1 provides a more detailed description of the sample.

Research Measures

Performance Measures. A complete description of performance criterion development work can be obtained from other Project A reports. This work included the development of the following measures: 1) Army-wide rating scales for evaluating soldiers in any first-tour Army job (Borman, Motowidlo, Rose, & Hanser, 1983; Pulakos & Borman, 1986); 2) Job-specific

Table 1

Description of the Sample

		Sex		Installation	
Army Job	N	Male	Female	CONUS[a]	Europe
Infantryman	673	673	0	568	105
Cannon Crewman	629	629	0	544	85
Armor Crewman	485	485	0	416	69
Radio Operator	351	300	51	329	22
Light-Wheel Vehicle Mechanic	618	577	41	508	110
Motor Transport Operator	659	598	61	554	105
Administrative Specialist	500	225	275	431	69
Medical Specialist	485	361	124	466	19
Military Police	680	629	51	552	128
Total:	5080	4477	603	4368	712

[a] CONUS = Continental United States

rating scales (Toquam, McHenry, Corpe, Rose, Lammlein, Kemery, Borman, Mendel, & Bosshardt, 1986); and 3) hands-on proficiency measures and job knowledge tests (Campbell, Campbell, Rumsey & Edwards, 1986). The Army-wide rating scales were constructed using the techniques of behaviorally-anchored rating scales (Smith & Kendall, 1963), and focus on performance dimensions relevant to any MOS (e.g., following rules, regulations, and orders; maintaining equipment). The job-specific scales were developed using the same approach; they focus on narrowly defined performance areas relevant to a specific job (e.g., loading cargo and transporting personnel, for motor transport operators). Finally, hands-on task proficiency measures tap skills in actually performing important tasks within a job, and the job knowledge measures contain paper-and pencil, multiple choice items assessing knowledge about how to perform the same important job tasks.

Work Environment Measures. The Army Work Environment Questionnaire (AWEQ), a revised 53 item multiple choice questionnaire was used to measure the following Army environmental constructs: 1) Resources/Tools/ Equipment, 2) Support, 3) Training/Work Assignment, 4) Job Importance, and 5) Cooperation/Cohesiveness. The Resources and Training constructs are job-oriented and the other three factors are more climate-related. The items on the AWEQ are answered using a 5-point frequency rating scale (e.g., 1 = Very Seldom or Never to 5 = Very Often or Always). Respondents were asked to indicate "how often" each environmental situation described in a questionnaire item occurs on their present job. For example, items consisted of statements such as "Important equipment changes or substitutions are made on your job without much advance notice," (Resources/Tools/Equipment), "You get recognition from supervisors for the work you do," (Support), and "You have the opportunity to practice or use the skills that are specific to your MOS" (Training). For the entire AWEQ, half of the items were worded negatively (as in the first example), and half positively (as in the remaining examples). Five standardized unit weighted factor scores are derived for the AWEQ. A more complete description of the scale development and field test results for the AWEQ can be found in Olson and Borman (1986).

Cognitive Ability. The Armed Services Vocational Aptitude Battery (ASVAB) is administered to all recruits prior to entering military service. The ASVAB, a general cognitive measure which contains 10 subtests, is used for making military selection and classification decisions. A composite measure of four ASVAB subtests, known as the Armed Forces Qualification Test (AFQT), was used as an assessment of general cognitive abilities in this research.

Temperament Measures. The Assessment of Background and Life Experiences (ABLE) inventory, which includes ten temperament/biodata scales, one biodata scale, and four response validity scales was administered as a self-report measure of soldier temperament in this research. The ten temperament scales are: 1) an 18-item, Emotional Stability scale, which assesses a person's characteristic affect and ability to react to stress; 2) a 20-item Non-delinquency scale, which measures how often a person violates laws, rules, or social norms; 3) an 11-item Traditional Values scale, that assesses how conventional, strict, or flexible a person's value system is; 4) a 15-item Conscientiousness scale, that measures respondents' degree of dependability, and tendencies to be organized and planful; 5) a 19 item, Work Orientation scale, that addresses how respondents feel about work and how they typically work; 6) a 12-item Self-Esteem scale, which measures how successful a person expects to be in life; 7) a 12-item Dominance scale, that measures respondents' tendencies to take charge and/or play a central and public role; 8) a 21-item Energy Level scale that assesses the degree to which one is alert, energetic, and enthusiastic; 9) a 16-item Internal Control scale which assesses both internal and external control as they pertain to achieving success on the job and in general life; and 10) an 18-item Cooperativeness scale that measures how easy it is to get along with the person providing the scale responses.

The Physical Condition scale is a 6-item biodata scale that measures to what extent respondents engage in physical activities such as sports and exercise. The four Response Validity scales (i.e., Non-random Response, Unlikely Virtues, Self-Knowledge, and Poor Impression) provide information on how soldiers have answered the ABLE. The major purpose of these validity scales is to determine the degree to which the respondents' answers are accurate. The sum of the weighted responses to each item on a scale serves as the scale score. All items have three options and receive weights of 1, 2, or 3.

From extensive field testing of the ABLE and principal factor analysis with varimax rotations emerged a three-factor solution for the temperament domain: 1) Surgency, 2) Dependability and 3) Adjustment constructs. The Surgency or Leadership/Achievement factor has items loading from the Self-Esteem, Work Orientation, Dominance and Energy-Level scales. The Dependability factor contains items from the Non-deliquency, Traditional

Values, Conscientiousness, Cooperativeness, and Internal Control scales. The Adjustment factor has items loading from the Emotional Stability scale. For purposes of this research, separate unit weighted factor scores were used as the temperament measures. For a more comprehensive discussion of the ABLE inventory, readers are referred to Hough (1984) and Hough and Ashworth (1986).

Procedures

The rating scales were administered to groups of 15 or fewer peers or supervisors of the target ratees after they were trained using a combination error and accuracy training program (e.g., Pulakos, 1984). On average, 1.90 supervisor raters and 3.26 peer raters per ratee provided these performance evaluations on the Army-wide and job-specific behavior-based rating scales. During the peer rating sessions, raters (who were in addition ratees and members of the research sample) also responded to the environmental questionnaire (AWEQ).

The ABLE inventory was administered to research participants in separate small group sessions. Hands-on task proficiency was measured by administering to each soldier in the sample 15 individual work samples representing 15 of the most important tasks for a designated job. Experienced job incumbents or supervisors were trained as scorers for the hands-on measures, and used an objective checklist to evaluate each soldier on the separate work sample tasks indicative of a specific job (Campbell, et al., 1986). Separate multiple-choice job knowledge tests for the nine Army jobs were administered to groups of 15-30 soldiers.

Analyses

Data analyses first examined the internal psychometric properties of the AWEQ scales, ABLE scales, the hands-on and job knowledge tests, and interrater reliabilities for the various rating instruments. Next, responses to the AWEQ were factor-analyzed in an effort to replicate the five-factor solution obtained for the environmental questionnaire during the extensive field tests. The ABLE responses were factor-analyzed to determine the underlying constructs of the temperament domain (Hough & Ashworth, 1986). When the performance rating domain was factor-analyzed an interpretable three-factor solution emerged: 1) Technical Skill and Effort, 2) Integrity and Control, and 3) Appearance (W. C. Borman, personal communication, April 1986). Composites derived from this solution were used in subsequent analyses, along with an overall effectiveness composite formed by unit weighting ratings on each dimension.

For purposes of the present research it seemed preferable to work with a single overall job knowledge test score and a single summary hands-on test score to represent these two performance areas for individual soldiers. Hence, a percent of items correct index was formed for the job knowledge area, and a percent of performance steps passed index served as the overall task proficiency score.

Correlational analyses were used to investigate the magnitude of the relationships between the work environment factors, temperament factors, ability, and the performance rating factors, as well as the job knowledge and hands-on tests. Multiple regression procedures were used to generate the path coefficients, that were in turn used to test relationships in a general model of influences on soldier performance.

Causal Ordering of the Variables in the General Model

The proposed general model of the influences on soldier performance displayed in Figure 1 expands the performance model developed by Hunter (1983) through: 1) including previously uninvestigated individual difference variables related to soldier temperament and perceptions of the work environment and 2) attempting to decompose the typical performance measurement domain into separate rating factors associated with a measure of overall soldier effectiveness.

Causal ordering of the performance variables (i.e., those variables from the job skills, ratings of dimensional effectiveness, and ratings of overall performance effectiveness domains) in the analysis was derived partly from the previous research of Hunter (1983) and Schmidt et al. (1986). Plausible relationships between individual differences and job conditions, perceptions of the work environment, and the performance domain are based on previously discussed theoretical work by Borman (1983; 1985) and Hough (1984), as well as on empirical research that focuses on environment-performance relationships (O'Connor et al. 1984; Olson & Borman, 1986; Steel & Mento, 1986).

These sources suggest a model (see Figure 1) in which influences on soldier performance work from relatively stable and consistent individual differences through less stable and more variable perceptions of the work

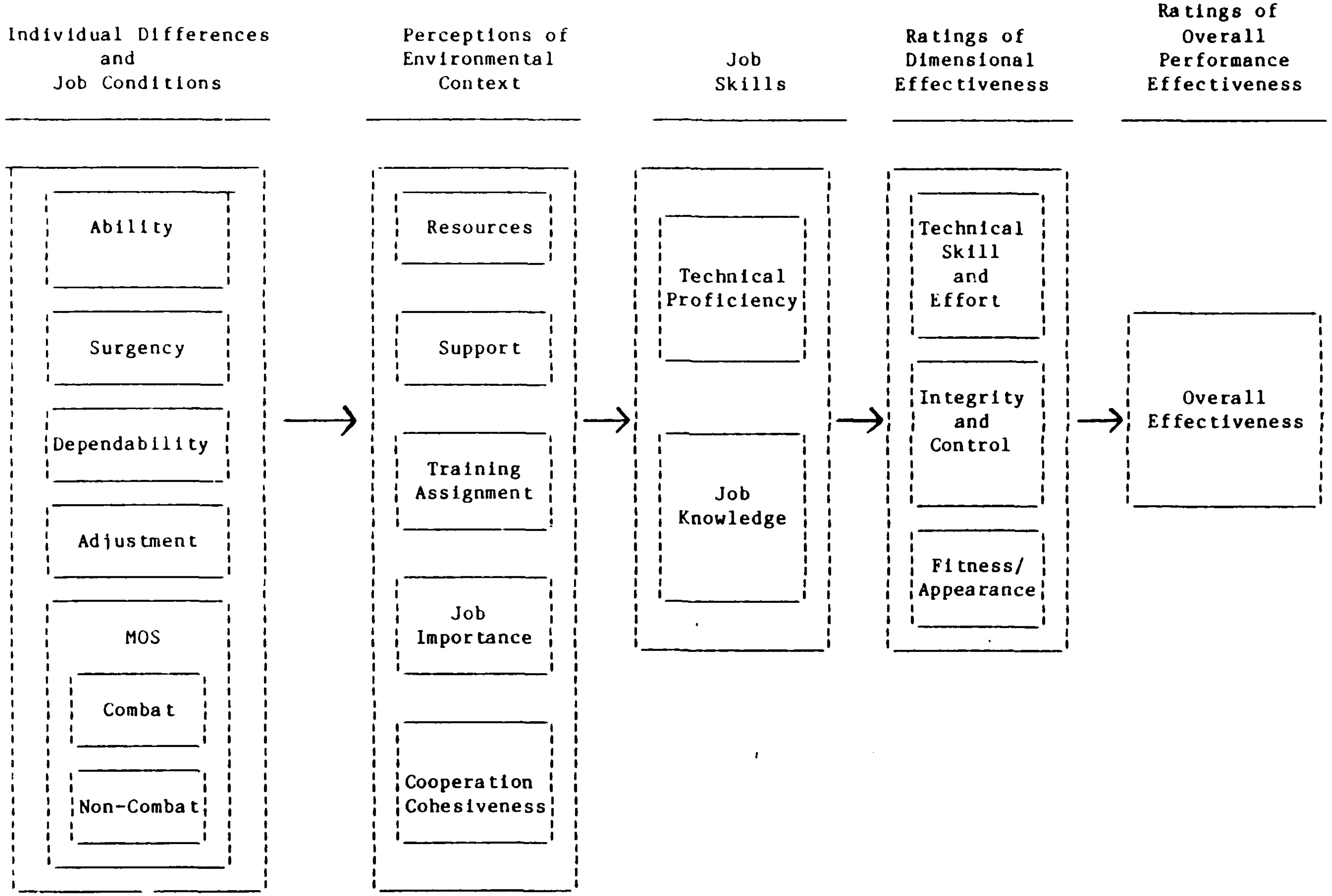


Figure 1. General Model: Influences on Soldier Performance

environment context to influence the broader performance domain. Ability, temperament, and type of job factors are viewed as independent or exogenous variables that may influence all endogenous variables, which are depicted to the right of them in the model.

Perceptions of the work en ronment context are endogenous factors that may affect job skills (e.g., job knowledge), ratings of dimensional effectiveness (e.g., Technical Skill and Effort), and overall soldier effectiveness. Environmental factors can have direct relationships with technical hands-on proficiency and job knowledge, and indirect effects on both the Army-wide ratings factors and overall effectiveness. Also, perceptions of the work environment may be affected by the individual temperaments soldiers bring to work settings. The job skills domain consists of performance variables that may directly affect ratings of dimensional effectiveness and indirectly affect overall effectiveness. The Army-wide rating factors (e.g., Integrity and Control) are also endogenous variables that may directly influence overall soldier effectiveness.

Since this general model assumes that a weak causal order exists among these variables and that the model is a causally closed system with respect to the influence of other explanatory variables, several possible relationships among the variables can be postulated.

First, it is proposed that strong direct relationships will be observed between cognitive ability and the more maximal task proficiency and job knowledge measures from the job skills domain. Second, ability will have indirect relationships with both ratings of dimensional effectiveness and ratings of overall soldier effectiveness through its direct effects on the measures (e.g., job knowledge) which represent the job skills domain. Third, the temperament variables (e.g., Surgency) will have direct relationships with the ratings of dimensional effectiveness (e.g., Integrity and Control and Appearance) and indirect associations (through the Army-wide rating factors) with ratings of overall soldier effectiveness. Also, temperament/dispositional tendencies and job variables may help determine perceptions of the work environment (particularly for the climate factors of Support, Job Importance, and Cooperation/Cohesiveness), which in turn may have indirect effects on performance ratings. Fourth, the more job-related environmental factors (Resources and Training Assignment) should have direct relationships with both technical proficiency hands-on measures and job knowledge tests, and indirect effects on the Army-wide rating factors related to Technical Skill and Effort. Fifth, the more climate-oriented environmental factors (i.e., Support, Job-Importance, and Cooperation/Cohesiveness) should be related to the Integrity and Control, and Appearance factors. Finally, within the performance domain, ratings of dimensional effectiveness will have direct relationships with ratings of overall soldier effectiveness.

RESULTS

Descriptive Statistics

Means, standard deviations, and the range of reliability estimates for the research measures are summarized in Table 2.

Table 2

Means, Standard Deviations, and Reliability Estimates of the Research Measures for the Total Sample

Research Measures	Means	Standard Deviations	Reliability[a] Estimates
Ability (AFQT Score)	51.99	20.40	.90 - .93
Temperament:			
Surgency	146.39	17.11	.80 - .85
Dependability	185.72	19.00	.80 - .85
Adjustment	41.00	5.75	.80 - .85
Work Environment:			
Resources	40.98	9.23	.80 - .87
Support	27.08	6.51	.78 - .84
Training	12.22	4.11	.71 - .87
Job Importance	20.72	4.22	.56 - .74
Cooperation/Cohesiveness	12.84	3.00	.64 - .75
Performance:			
Overall Effectiveness	4.60	.84	.50
Task Proficiency (Hands-On)	70.67	10.60	.52 - .79
Job Knowledge	61.17	10.58	.82 - .89
Army-wide Rating Factors:			
Technical Skill and Effort	4.38	.79	.50 - .60
Integrity and Control	4.59	.86	.50 - .60
Appearance	4.86	.89	.50 - .60

[a] Work environment reliabilities are coefficient alphas (measures of internal consistency); Temperament reliabilities are test-retest; Hands-on and Job Knowledge reliabilities are split-half indices; Interrater reliabilities are shown for the Army-wide rating factors.

Reliability estimates presented for the AWEQ are measures of internal consistency (coefficient alphas). For the environmental factors most of these reliabilities are reasonably high. Similiarly, the test-retest reliabilities (interval of two weeks to two months) are .80 or above and quite high for the temperament factors. The interrater reliabilities for the Army-wide rating factors range from .50 to .60. For the maximal performance measures, the split-half reliabilities of the job knowledge measures were higher than those observed for the hands-on task proficiency measures.

Table 3 presents the intercorrelation matrix for the research measures. For the predictor measures, low intercorrelations were found between ability and both temperament ($\underline{r}$s ranged from .08 to .16) and environment ($\underline{r}$s ranged from -.05 to .06) factors. Within the temperament domain, Surgency was highly intercorrelated with Dependability and Adjustment. For the work environment predictors, Support was strongly associated with Job Importance ($\underline{r}$ = .45), Resources ($\underline{r}$ = .46) and Cooperation/Cohesiveness ($\underline{r}$ = .44). Correlations between the temperament and work environment factors showed strong relationships between Surgency and Dependability and the work environment constructs of Support and Job Importance.

In summary, these patterns of intercorrelations among the predictors suggest appropriately that cognitive ability, as measured by the AFQT, is relatively distinct from temperament and work environment factors. In

Table 3

Intercorrelations Among the Research Measures for the Total Sample

Predictors	1	2	3	4	5	6	7	8	9
1. Ability									
Temperament:									
2. Surgency	13								
3. Dependability	08	63							
4. Adjustment	16	57	44						
Environment:									
5. Resources	-05	06	21	14					
6. Support	03	20	28	15	46				
7. Training	-04	02	05	04	25	24			
8. Job Importance	0	27	30	15	35	45	24		
9. Cooperation/Cohesiveness	06	13	19	16	27	44	19	24	

Performance Criteria	1	2	3	4	5	6
1. Job Knowledge						
2. Task Proficiency (Hands-On)	36					
3. Overall Effectiveness	21	13				
Army-Wide BARS Rating Factors:						
4. Technical Skill and Effort	25	16	86			
5. Integrity and Control	18	06	74	73		
6. Appearance	02	02	63	60	51	

Note. Correlations greater than .03 are significant at $p < .01$.

comparison, the Surgency factor, which assesses leadership/achievement, is not only highly intercorrelated with the other temperament factors but has relationships with the work environment domain.

Table 3 also shows the correlation matrix for the performance measures. Generally, the maximal (e.g., job knowledge) and typical (e.g., rating factors) performance criteria are not highly related. For the maximal performance measures, a correlation of .36 was noted between hands-on task proficiency and job knowledge tests. In contrast for the typical measures, the Army-wide rating factors tended to be highly intercorrelated. Further, a considerable portion of the variability in overall soldier effectiveness was attributed to Technical Skill and Effort (74%), Integrity and Control (55%), and Appearance (40%), which are the Army-wide rating factors.

Correlations Between Predictors and Performance Measures

Table 4 presents the relationships between ability, temperament, work environment and the performance criteria. As was predicted, ability was not highly correlated with performance ratings, but had a strong relationship with the job knowledge or "can do" type of performance measure. The temperament factors tended to have stronger correlations with the Army-wide rating factors than did the environmental variables. Specifically, Surgency had significant correlations in the mid to high 20's with Technical Skill and Effort, Appearance, and Overall Soldier Effectiveness. The highest significant correlation ($\underline{r}$ = .31) was found between the temperament factor of Dependability and the Army-wide rating factor of Integrity and Control.

Table 4

Correlations Between Predictors and Performance Measures

	Performance Measures[a]					
Predictors	1	2	3	4	5	6
Ability	.10	.42	.14	.10	-.05	.11
Temperament						
Surgency	.02	.08	.28	.15	.25	.26
Dependability	-.04	.11	.25	.31	.24	.26
Adjustment	.03	.12	.16	.12	.15	.16
Environment						
Resources	-.06	-.07	.05	.12	.14	.06
Support	0	.05	.17	.24	.18	.19
Training	.23	.06	.07	.06	.06	.06
Job Importance	.06	.07	.20	.20	.16	.20
Cooperation/Cohesiveness	.05	.09	.09	.12	.10	.11

Note. *N* ranges from 4274 to 5035.

[a] Performance measures are: 1 = Hands-on Test (Average % go for all tasks, 2 = Job Knowledge (Average % for all tasks), 3 = Army-wide BARS Technical Skill and Effort factor, 4 = Army-wide BARS Integrity and Control factor, 5 = Army-wide BARS Appearance factor, and 6 = Overall Effectiveness.

Correlations greater than .05 are significant at $p < .001$.

The temperament factors were not significantly associated with any of the maximal performance measures (e.g., job knowledge). In contrast to this finding, a significant correlation ($\underline{r}$ = .23) was observed between the work environment factor of Training and hands-on task proficiency. The Job Importance and Support environmental factors were consistently related to the performance ratings (correlations of practical and statistical significance ranged from .16 to .24). Although correlations between the set of performance criteria and the environmental factors related to Resources and Cooperation/Cohesiveness were significant, their average absolute value was low, respectively .08 and .09.

Multiple Regression Analysis

A series of ordinary least-squares multiple regression analyses were conducted to 1) determine the amount of variability in each dependent or endogenous variable that could be attributed to a specified set of variables from the general model (See Figure 1) and 2) generate standardized beta weights for use in the subsequent path analysis.

In these multiple regression analyses, five factor composites represented the work environment constructs, three unit weighted composites represented the temperament domain, and three summary composites and an overall effectiveness index reflected soldier effectiveness on constructs relevant to any Army job (e.g., Integrity and Control). Single summary measures of job knowledge and task proficiency represented performance, respectively, in those two criterion domains.

Army jobs were grouped into combat and non-combat for these multiple regression analyses. Combat jobs include Infantryman, Cannon Crewman and Armor Crewman. The remaining six Army jobs were classified as non-combat.

For purposes of these analyses, each endogenous variable in the general model was regressed on all variables presumed to be antecedent to it and then the standardized regression coefficients were used to estimate path coefficients tested in the general model. After multiple regression techniques were applied to the general model, the magnitude of the standardized beta coefficients were examined. Based on considerations related to both meaningfulness of results and statistical significance, paths with standardized beta weights of less than .15 were deleted (i.e., path was considered zero) and multiple regression analyses were rerun on the restricted model (i.e., model with the decreased number of variables) to determine the path coefficients. The .15 criterion was considered a stringent test for inclusion of variables in the restricted model.

The beta weights which were used to estimate the path coefficients are summarized in Table 5. Table 6 compares the variance explained by the general and restricted models. As shown in the restricted model, 78% of the variability in overall effectiveness can be attributed to the Army-wide rating factors. This percentage of explained variance is maintained even after 12 variables are trimmed from the general model. The variability in the separate Army-wide rating factors is accounted for by different sets of variables. Specifically, the explained variance (12%) in the Technical Skill and Effort factor can be attributed to Surgency and job knowledge. An identical amount of explained variance (12%) in the Integrity and Control rating factor was accounted for by the

Table 5

Standardized Beta Weights from Multiple Regression Analyses Used to Estimate Path Coefficients in the General Model

Independent Variables	Dependent Variables										
	1	2	3	4	5	6	7	8	9	10	11
1. Technical Skill and Effort	.59**										
2. Integrity and Control	.21**										
3. Appearance	.17**										
4. Technical Proficiency	.00	.10**	.03	.04*							
5. Job Knowledge	.01	.15**	.10**	-.01							
6. Resources	-.02*	-.04*	-.03	.05*	-.13**	-.13**					
7. Support	.01	.10**	.15**	.09**	-.04	-.01					
8. Training	-.00	.00	-.02	.01	.26**	.09**					
9. Job Importance	.01	.10**	.08**	.01	.07**	.04*					
10. Cooperation/ Cohesiveness	.01	-.02	-.01	.01	.04*	.06**					
11. Ability	.00	.04*	.04*	-.07**	.11**	.40**	-.08**	.00	-.06**	-.05*	.03
12. Surgency	.01	.19**	-.09**	.18**	.02	-.07**	-.18**	.02	-.03	.16**	-.02
13. Dependability	-.00	.07**	.30**	.09**	-.08**	.12**	.26**	.26**	.05*	.20**	.14**
14. Adjustment	.01	-.02	-.01	.01	.05*	.04*	.14**	.03	.04*	-.02	.10**
15. Combat/ Noncombat	-.01	-.00	-.02	-.01	-.03	.04*	-.11**	-.02	-.01	-.10**	-.10**

Note. The dependent variables are 1 = Overall Effectiveness, 2 = Technical Skill & Effort, 3 = Integrity & Control, 4 = Appearance, 5 = Technical Proficiency, 6 = Job Knowledge, 7 = Resources, 8 = Support, 9 = Training, 10 = Job Importance, 11 = Cooperation/Cohesiveness.

Table 6

Variance Explained by General and Restricted Models

	General Model		Restricted Model	
Dependent Variables	R^2	Number of Predictors	R^2	Number of Predictors
Overall Effectiveness	.78	15	.78	3
Technical Skill & Effort	.16	12	.12	2
Integrity & Control	.15	12	.12	2
Appearance	.11	12	.06	1
Technical Proficiency	.09	10	.05	1
Job Knowledge	.20	10	.17	1
Training	-.01	5	.00	0
Support	.09	5	.08	1
Resources	.08	5	.05	2
Job Importance	.11	5	.10	2
Cooperation/Cohesiveness	.05	5	.00	0

Note. Goodness of Fit Index = Q = .70 for the model.

individual soldier characteristic of dependability and soldier perceptions of the level of support received from significant others in the work environment. Six percent of the explained variance in the Appearance rating factor was accounted for by Surgency, which is a more leader and achievement-oriented soldier temperament.

In the maximal performance domain, some of the variance (5%) in technical proficiency (hands-on test performance) was accounted for by soldier perceptions of training opportunities and work assignments (environmental factor). Cognitive ability (beta weight = .40) accounted for a large portion of the explained variance (17%) in scores on job knowledge tests.

Three of the five environmental constructs were retained in the restricted model. These included Resources, Support and Job Importance factors. A small portion of the variability (5%) in perceptions of resources was significantly associated with the soldier temperament factors (Dependability, beta weight = .26; Adjustment, beta weight = .15; and Surgency, beta weight = -.18). A significant portion of the total variability (8%) in soldier perceptions of support was explained by Dependability. Further, perceptions of job importance were attributed primarily to the temperament factors of Surgency and Dependability.

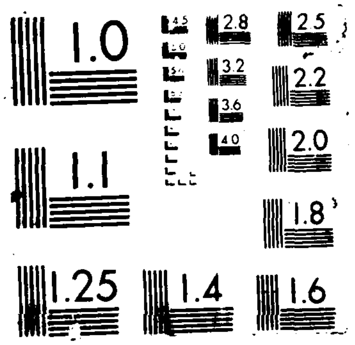
In the work environment domain, no predictors were retained in the restricted model that accounted for significant variance in soldier perceptions of Training or Cooperation/Cohesiveness.

Path Analysis

The results of the exploratory path analysis are presented in Figure 2. The standardized beta weights generated from the multiple regression analyses function as the path coefficients and indicate the relative strength of the variables retained in the restricted model. This results in an overidentified model with a goodness of fit index of Q = .70 (See Pedhazur, 1982, p. 621 for a discussion of this index). As Q appraoches 1.0 the fit of the data to the model becomes maximal.

The path analysis indicates that cognitive ability (as measured by AFQT) is the main determinant of jol knowledge. In our restricted model with its very stringent criterion (.15) for inclusion of variables, ability was not found to affect technical proficiency. However, when this criterion is changed to .10 for acceptance of standardized beta weights into the restricted path model, a significant path coefficient of .11 is obtained between ability and hands-on technical proficiency. This path coefficient of .11 between ability (AFQT) and hands-on measures supports other military research (Schmidt, Hunter, & Outerbridge, 1986) which found a coefficient of .13 between AFQT and work sample performance. Hence, this finding suggests that the reason no direct effect of ability on technical proficiency was observed in our research can be linked to the

AD-A196 274 IMPROVING THE SELECTION CLASSIFICATION AND UTILIZATION 4/5
OF ARMY ENLISTED P (U) HUMAN RESOURCES RESEARCH
ORGANIZATION ALEXANDRIA VA J P CAMPBELL MAY 88
UNCLASSIFIED HUMRRO-IR-PRD-87-12-SUPPL F/G 5/9 NL


1.0
2.8
2.5
3.2
2.2
3.6
4.0
2.0
1.1
1.8
1.25
1.4
1.6

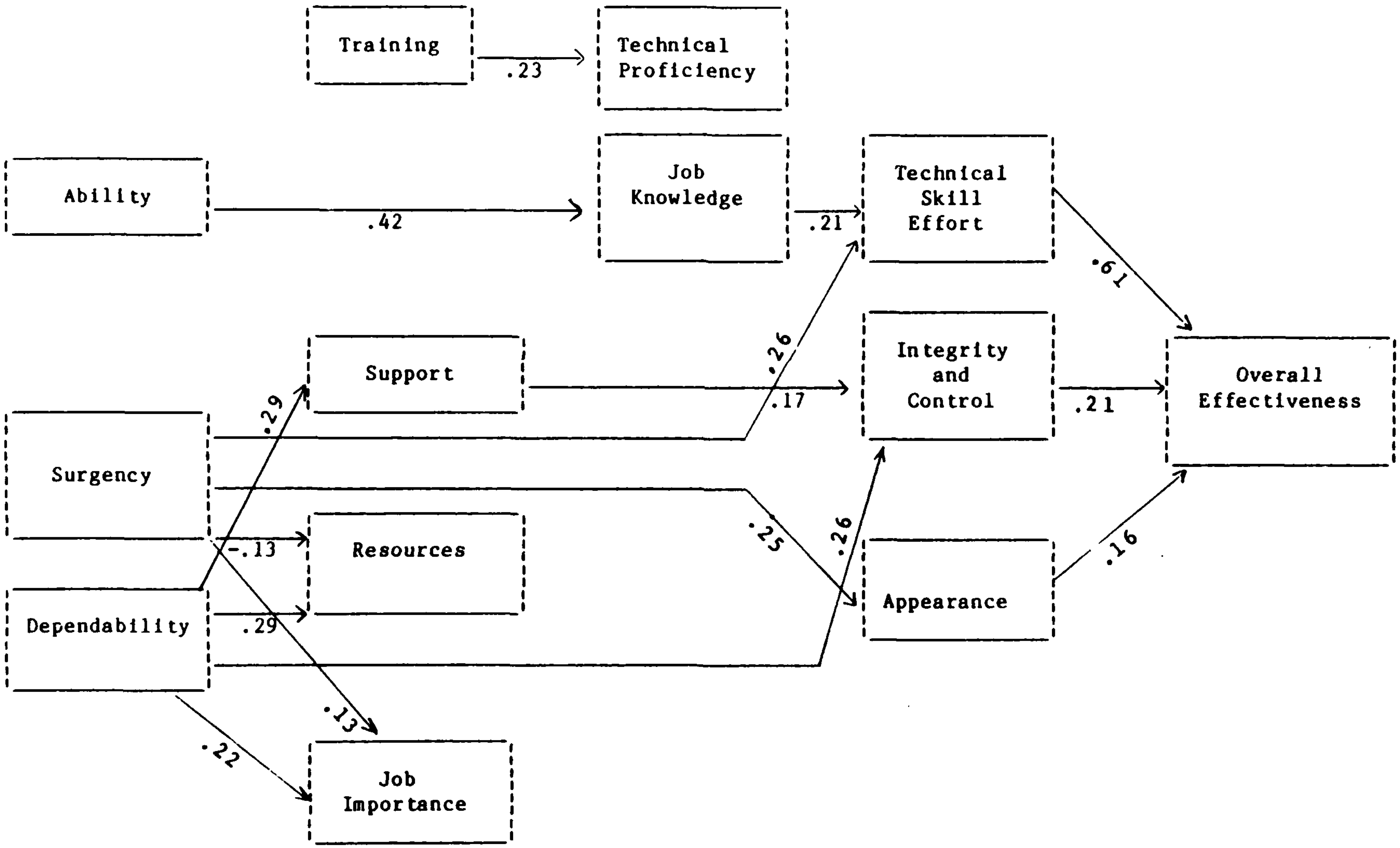


Figure 2. Path Analysis of the Restricted Model

assumptions of our model as opposed to an actual lack of relationship between these variables. Ability had an indirect effect on overall soldier effectiveness through its impact on job knowledge and Technical Skill and Effort. Overall soldier effectiveness is directly affected by the Army-wide rating factors, particularly Technical Skill and Effort (path coefficient = .61).

The ratings of dimensional effectiveness were impacted on by job skills, environmental factors, and temperament variables. Specifically, job knowledge and Surgency had direct associations with Technical Skill and Effort. Soldier perceptions of environmental support directly affected the Integrity and Control rating factor. The Appearance Army-wide rating factor was solely determined by soldier temperament associated with Surgency (leader- and achievement-oriented disposition).

In the job skills domain, hands-on technical proficiency was directly related to soldier perceptions of training opportunities and work assignment. Findings from this analysis suggest that the differences in soldier temperament and dispositional tendencies directly affect soldier perceptions of the work environment. For example, soldier perceptions of the level of supervisory support was related to their having a dependable disposition, which was characterized by non-delinquency, traditional values, cooperativeness, and conscientiousness. Further, soldier perceptions of available environmental resources and the importance of their jobs was directly impacted on by their having temperaments which reflected high self-esteem, dominant disposition, high energy level and strong work orientation (Surgency factor) as well as dependability.

Several indirect relationships were found between factors from the individual differences domain and overall soldier effectiveness. In particular, surgency was associated with overall effectiveness through its separate direct effects on Technical Skill and Effort and Appearance. Soldier temperament characterized by dependability impacted indirectly on overall effectiveness through its effect on soldier perceptions of environmental support and the Army-wide rating factor of Integrity and Control.

DISCUSSION

The restricted model developed through path analysis suggests that cognitive ability has a direct effect on job knowledge and indirect effects on dimensional effectiveness and overall effectiveness domains. Cognitive ability is the only determinant of job knowledge. Further, cognitive ability had indirect effects on Technical Skill, and Effort, and overall soldier effectiveness through its direct relationship with job knowledge.

Only the Surgency and Dependability factors from the temperament domain were retained in the restricted model. Surgency related directly to the work environment variables and Army-wide ratings of dimensional effectiveness, and indirectly on overall effectiveness. The Technical Skill and Effort rating factor was influenced by soldier dispositions characterized by high achievement needs, strong work orientation, and leadership capabilities (Surgency factor). Also, Surgency had a direct association with the Appearance rating factor. The Dependability temperament construct, which reflects soldiers' adherence to laws/norms, traditional values, and tendencies to be organized and planful (internal control) was directly related to the Army-wide rating factor of Integrity and Control, which denotes following military rules and displaying self-control. These findings for the relationships between the temperament factors and the Army-wide rating factors make conceptual sense, and tentatively suggest that soldiers with certain dispositions may have an easier time adapting to military life. Further, this finding supports previous military research (cited in Hough, 1984), which found that temperament constructs related to Dependability predicted military adjustment criteria.

Several indirect relationships were observed between temperament factors and performance ratings. For example, dependable, conscientious soldiers (Dependability factor) tended to perceive the Army work environment as supportive (Support factor) and this in turn was related to soldiers' ability to obey orders, display appropriate respect for the military chain-of-command, and exhibit self-control (Integrity and Control), and consequently affected ratings of overall effectiveness provided for these soldiers.

Temperament factors also had direct impacts on soldier perceptions of the work environment. For instance, soldiers who came to the Army with dependable dispositions (non-delinquent, cooperative, and conscientious) tended to view their job as worthwhile, held perceptions of the chain-of-command as supportive and responsive to their needs, and believe that sufficient resources in terms of tools, personnel and equipment were available to effectively complete their job assignments.

One of the most interesting work environment findings was related to the impact of soldier perceptions of their training and work assignment on hands-on technical proficiency. The ability to perform the tasks of a job under standardized conditions was associated with soldiers' perceptions of their opportunities to receive training in their MOS as well as practice the new skills acquired in training; and less affected by the abilities soldiers brought to the job. For purposes of this research, it is important to note that cognitive abilities brought to the job are measured by AFQT. When the Aptitude Area Composites from the Armed Services Vocational Aptitude Battery (ASVAB) are used as ability measures, findings from other Project A data suggest that a stronger relationship ($\underline{r}$ = .36) exists between ability and hands-on performance (HumRRO, AIR, PDRI, & ARI, 1986).

Of the four environmental constructs retained in the path analysis, only perceptions of Training and Support impacted on the performance domain. These relationships between work environment and performance components provide support for our conceptualization of the Army work environment in terms of job-and climate-oriented constructs. Specifically, Training, which was conceived of as a job-oriented environmental factor related appropriately to the "can do" or job skills component of performance. In contrast, the more climate-oriented Support factor impacted on the "will do" or affective/motivational aspects of the performance space (i.e., Integrity and Control).

Results from this empirical research support the ideas advanced by Borman, Motowidlo, Rose, and Hanser (1985) in their Model of Soldier Effectiveness, which proposed that "being a good soldier" or "having overall worth to the Army" requires more than just performing job tasks competently. Specifically, these findings demonstrate the importance of considering non-cognitive (temperament) variables and soldiers' perceptions of the work environment as determinants of both dimensions of soldier effectiveness and maximal performance measures. Further, these efforts aimed at dimensionalizing the performance domain in terms of both job skill components and other more motivational/social/adjustment elements of the criterion space, take an important step toward "defining performance domains and devising valid and fair measures of them", as advocated by Hakel (1986).

Several variables proposed in the general model did not affect performance or have links to other variables in the path analysis. Actual job assignment, combat or non-combat MOS, did not affect, for instance, soldier perceptions of the importance of their jobs or subsequent Army-wide rating factors (e.g., Appearance and Fitness). One could assume that serving in a combat job might be related to good physical conditioning. Soldier temperaments characterized by adjustment and ability to cope with stress did not directly or indirectly affect Army-wide rating dimensions. It might be expected that soldier temperaments that reflected more effective mechanisms for handling stress would have linkages to the Integrity and Control dimension, particularly as it related to exhibiting self-control during stressful personal or financial crises.

In summary, this research has proposed a first approximation to a model of influences on soldier performance, which examines individual differences in ability and soldier temperament, job conditions/assignment, perceptions of the work environment, and measures of both typical and maximal performance. Several limitations of the current research should be noted. First, the causal flow proposed in our recursive model is unidirectional, which does not consider the possibility of reciprocal causation between variables in the model. For example, our model does not consider that the importance of a job to which a soldier is assigned may directly affect on what he/she expects to achieve on that job. Further, this model did not test whether job importance indirectly affects Technical Skill and Effort through its impact on a soldier's dispositional tendencies to be achievement-oriented.

Second, it is assumed that all relevant variables are included in the model that was tested. Finally, it was assumed that our variables were measured without error.

Generally, there was only a small decrease in explanatory power across dependent criteria between the general and restricted models of influences on soldier performance. For example, two predictors accounted for 12% of the variance in the Control and Integrity rating factor under the restricted model as compared to 12 predictors explaining only 15% of the variability in the general model. Although a fair fit (Q = .70, Goodness of Fit index) was obtained between our data and the model, more research is needed to further clarify the process through which personal characteristics and environmental perceptions influence job performance criteria. Specifically, future research should consider latent variable models with multiple indicators of complex constructs. Rather than the exploratory analysis used in this research, subsequent work should apply confirmatory analytic techniques (e.g., LISREL), that better address unmeasured variable problems, reciprocal causation, and measurement error.

REFERENCES

Bandura, A. (1978). The self system in reciprocal determinism. *American Psychologist*, *3*, 344-358.

Borman, W. C. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. Landy, S . Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory*. Hillsdale, NJ: Lawrence Erlbaum Associates .

Borman, W. C., Motowidlo, S. M., Rose, S. R., & Hanser, L. M. (1984, August). A model of individual performance effectiveness: Thoughts about expanding the criterion space. *Integrated Criterion Measurement for Large-Scale Computerized Selection and Classification*. Symposium conducted at the 92nd A nnual convention of the American Psychological Association, Toronto, Canada.

Campbell, C. H., Campbell, R. C., Rumsey, M.G., & Edwards, D. C. (1985). *Development and field test of Project A task-based, MOS-specific criterion measures*. Alexandria, VA: Human Resources Research Organization.

Campbell, J. P., & Pritchard, R. D. (1976). Motivation theory in industrial and organizational psychology. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 63-130). Chicago, IL: Rand McNally.

Dunnette, M. D. (1976). Aptitude, abilities, and skills. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 473-520). Chicago, IL: Rand McNally.

Frederiksen, N., Jensen, O., & Beaton, A. (1977). Prediction of organizational behavior. New York: Pergamon Press.

Guion, R. M. (1983). Comments on Hunter. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hakel, M. D. (1986). Personnel selection and placement. Annual Review of Psychology, 37, 351-380.

Hough, L. M. (1984). Identification and development of temperament and interest constructs and inventories for predicting job performance of Army enlisted personnel. Minneapolis, MN: Personnel Decisions Research Institute.

Hough, L. M., & Ashworth, S. (1986). Project A concurrent validity data analyses: Temperament and interest predictors. Personnel Decisions Research Institute, Minneapolis, MN.

HumRRO, AIR, PDRI,& ARI (1986). Project A: Preliminary Incremental Validities. Personal Communication.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisory ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory (pp. 257-266). Hillsdale, NJ: Lawrence Erlbaum Associates.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. Journal of Applied Psychology, 67 (2), 219-229.

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. Psychological Bulletin, 81, 1096-1112.

James, L. R., Hater, J. J., Gent, M. J. & Bruni, J. R. (1978). Psychological climate: Implications from cognitive social learning theory and interactional psychology. Personnel Psychology, 31, 783-814.

Magnusson, J. (1981). Toward a psychology of situations: An interactional perspective. Hillsdale, NJ: Lawrence Erlbaum Associates.

O'Connor, E. J., Peters, L. H., Pooyan, A., Weekley, J., Frank, B., & Erenkrantz, B. (1984). Situational constraint effects on performance, affective reactions and turnover: A field replication and extension. Journal of Applied Psychology, 69 (4), 663-672.

Olson, D. M., & Borman, W. C. (1986). Development and field tests of the Army Work Environment Questionnaire (Working Paper RS-WP-86-06). Alexandria: U. S. Army Research Institute.

Pedhazur, E. J. (1982). Multiple regression in behavioral research. New York: Holt, Rinehart & Winston.

Peters, L. H., & O'Connor, E. J. (1980). Situational and work outcomes: The influences of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Peters, L. H., & O'Connor, E. J., & Rudolf, C. J. (1980). The behavioral and affecting consequences of performance-relevant situation variables. Organizational Behavior and Human Performance, 25,79-96.

Petersen, N. (Ed.) (1986). Report on the development of the trial battery for Project A. Research Product 86-4. Prepared by Personnel Decisions Research Institute for the U. S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*, 581-588.

Pulakos, E. D., & Schmitt, N. (1983). A longitudinal study of a valence model approach for the prediction of job satisfaction of new employees. *Journal of Applied Psychology*, *68* (2), 307-312.

Pulakos, E. D., & Borman, W. C. (Eds.) (1986). *Development and field test of Army-wide rating scales and the rater orientation and training program*. Alexandria: Human Resources Research Organization.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability of job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71* (3), 432-439.

Schneider, B. (1975). Organizational climate: An essay. *Personnel Psychology*, *28*, 447-479.

Schneider, B. (1983). Work climates: An interactionist perspective. In N. W. Feimer and E. S. Geller (Eds.), Environmental psychology: *Directions and perspectives*. New York: Praeger Press.

Schneider, B., & Reichers, A. E. (1983). On the etiology of climate. *Personnel Psychology*, *36*, 19-39.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for ratings scales. *Journal of Applied Psychology*, *47*, 149-155.

Staw, B. M., & Ross, J. (1985). Stability in the midst of change: A dispositional approach to job attitudes. *Journal of Applied Psychology*, *70* (3), 469-480.

Steel, R. P., & Mento, A. J. (1986). Impact of situational constraints on subjective and objective criteria of managerial job performance. Organizational Behavior and Human Decision Processes, 37, 254-265.

Toquam, J. L., McHenry, J. J., Corpe, V.A., Rose, S. R., Lammelin, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1985). Behaviorally-anchored rating scales for nine MOS: Development activities and field test results. Minneapolis, MN: Personnel Decisions Research Institute.

Watson, T. W., O'Connor, E. J., Eulburg, J. R., & Peters, L. H. (1983, October). Measurement and assessment of situational constraints in Air Force work environments: A brief summary. Proceedings of the 25th Annual Conference of the Military Testing Association.

WEIGHTING PERFORMANCE CONSTRUCTS IN COMPOSITE MEASURES OF JOB PERFORMANCE

Robert Sadacca Maria Veronica deVera Ani DiFazio

Human Resoures Research Organization

Leonard A. White

U.S. Army Research Institute

Presented on symposium,
"Multiple Criteria and Multiple Jobs: Will One Model Fit All?"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

Introduction

Performance evaluation is an important consideration in every type of organization. A good and accurate evaluation system improves the utilization of resources by fostering improvements in work performance, improving job assignment and assigning work more efficiently (Bernardin and Beatty, 1984). Accurate assessment of performance requires careful evaluation of the job's multidimensional tasks. The importance of each dimension to job success is generally not the same across jobs. The importance of each dimension should, therefore, be determined when evaluating overall job performance. The present study is concerned with the application of the principles of accurate performance evaluation in the Army's family of Military Occupational Specialties (MOS). The Army has the same basic concerns: A desire to improve their selection, classification and utilization of enlisted personnel. The development of measures of overall job performance for each MOS is a critical step in achieving this goal.

An appropriate procedure should be used in assessing the importance of each performance dimension in the measurement of overall job success. There have been several methods that have been recommended for assigning weights to performance dimensions in such a way that it reflects the factors' relative importance to overall performance. These four procedures have been emphasized in the literature: 1) The Two Factor-At-A-Time Conjoint procedure; 2) The Full-Approach Conjoint procedure; 3) The Kelly Bids system; and 4) The Kane method.

In a conjoint procedure the repondents are asked to rank order, rate, or otherwise react to one or more sets of profile descriptions which vary along the dimensions of interest. The relative weights for the dimensions can be inferred from the relationships between the dimension variance built into the descriptions and the rank orders or ratings (the dependent variable) given the profiles. Users of this type of methodology have generally emphasized predictive validity and regarded explanation (in terms of relative weights) as a desirable (but secondary) objective (Green and Srinivasan 1978). The Two-Factor-At-A-Time and the Full-Profile approaches have been generally used in conjoint procedures.

The Two-Factor-At-A-Time is also referred to as the Trade-off procedure (Johnson, 1974). In this procedure the performance factors are evaluated on a two-at-a-time basis. The evaluators are usually asked to rank the various combinations of each pair of factor levels from most preferred to least preferred (Green and Srinivasan, 1978). The advantages of using this procedure are that it is simple, reduces information overload, and lends itself to mail questionnaire administration. There are, however, a few limitations. This procedure has been criticized as being unrealistic because there are other factors that must also be taken into consideration in the overall evaluation. Some researchers (Johnson and Vandyke, 1975; Green, 1974) have pointed out that the total number of required evaluations is quite large when there there are multiple levels within the dimensions. In these circumstances the respondents may attend to one dimension first before considering the other (Johnson, 1976).

The Full-Profile approach attempts to address some of the limitations of the Two-Factor-At-A-Time procedure. This approach follows the same procedure as the former approach but utilizes the complete set of factors in the descriptions. It gives a more realistic description of the stimuli being judged by defining the levels on all of the factors and possibly taking into account the potential environmental correlations between the factors in real stimuli (Green and Srinivasan, 1978). It is, however, not devoid of limitations. The possibility of information overload is highly likely as the number of factors in the profile increases. Furthermore, the respondents may simplify the task by ignoring variations in the less important factors or by simplifying the factor levels themselves (Green and Srinivasan, 1978). Therefore, this procedure is generally limited to five or six factors.

The type of presentation used for these two procedures are verbal descriptions, paragraph descriptions, and pictorial representation. The Two-Factor has primarily used verbal descriptions. The appropriate type of presentation will, however, differ depending on the type of factors being considered in the study.

The measurement scale used for these conjoint procedures is either nonmetric (paired comparisons, rank order), or metric (rating scales assuming interval scales, ratio scales obtained by constant-sum paired comparisons). For the Two-Factor, the nonmetric scale is more appropriate because the rank order of the cells in a tradeoff table need not depend on the levels of the missing factors, except if the attributes are correlated (Green and Srinivasan, 1978). The metric scale, however, provides increased information and lends itself to administration by mail.

The effectiveness of these two procedures have been ev luated by several researchers. Montgomery, Wittink, and Glaze (1977) reported that the Two-Factor procedure yielded higher predictive validity. Their study focused on job choices made by MBAs. The study had a total of eight attributes. Alpert, Betak, and Golden (1978) in their study of commuters' choice of transportation modes reported that the goodness-of-fit to input data was better for the Two-Factor. A total of nine attributes were used in their study. Jain, Acito, Malhotra, and Mahajan (1978), on the other hand, reported that the two methods yielded approximately the same level of cross-validity in the context of choosing checking accounts offered by various banks. Five attributes were used in this study. Oppedijk van Veen and Beazley (1977) found that the utilities determined by the two methods were roughly similar in the context of a durable good product class when using three attributes.

Another procedure also used for weighting purposes is the Kelly Bids System. In this procedure the respondents are asked to allocate 100 points across the criterion dimensions on the basis of their relative importance. An average is then compiled across the ratings of the respondents. Schmidt (1977) found this procedure better than others because the focus is on the hypothetical "true" criterion. However, this method gives no consideration to the extent to which the various dimensions can actually be measured (Bernardin and Beatty, 1984).

Kane (1980) maintained that observability and uncertainty should also be considered critical in all appraisal situations. He, therefore, proposed the

Kane method for assigning weights to performance factors. An important aspect to this procedure is the designation of a level of specificity for assigning importance weights (e.g., task level) prior to any activity. The respondents are then asked to identify the component having the least importance for measuring overall effectiveness. This component is assigned a weight of 1.0. The respondents are then asked to compare the remaining factors to the least important component. They are to assign weights to the remaining factors such that the weights reflect how many times more important each factor is compared to the least important factor. Both the Kelly Bids method and the Kane method are adaptable to the different purposes of appraisal (Bernardin and Beatty, 1984).

All four procedures for assigning weights to performance factors have been shown to work well in a variety of settings. The appropriateness of the methodology depends to a great extent on the purposes and the type of factors and variables of the research endeavor.

The present report focuses on the research that was conducted to determine the best method for obtaining importance judgments regarding how to weight performance construct scores to form an overall composite index of performance for the 19 MOS comprising the Project A sample of jobs. After selecting two of the methods on the basis of the experimental results and other considerations, construct weighting data were gathered from noncommissioned officers (NCOs) and officers familiar with each MOS. The results of preliminary analyses of that data are presented at the end of the report. (At the time this report was prepared data were still being collected.)

Method and Results of Initial Weighting Method Experiments

Three field experiments were conducted to select the construct weighting procedures. These procedures were used to obtain the subject matter expert (SME) judgments concerning the relative importance of the construct weights for each MOS. The primary focus of the experiments was on the weighting procedures themselves and not on the weights of the constructs for given MOS. Our interest in conducting the experiments was in selecting one or more construct weighting procedures that would be acceptable to the Army and would yield a reliable, valid set of weights for each of the sampled MOS when tne procedures were applied by the SME. The three experiments were related in the sense that the weighting procedure selected as a result of the first experiment was also used in the second and third experiments to further evaluate that and other procedures. The experiments and their results will be described briefly prior to describing how the MOS construct weights were obtained.

Experiment 1 - Procedure and Results

Sixteen Army officers participated in the first experiment. The officers were stationed at Ft. Mead, MD, and Ft. Monroe, VA. Their task was to assign relative weights to 6 performance constructs for three MOS, Infantryman, Wheel Vehicle Repairer, and Administrative Specialist. At the time the experiment was conducted in Summer, 1985, the Project A performance constructs had not been selected. Therefore, a set of 6 constructs, whose weights might be expected to vary considerably was put together by the experimenters. The six performance constructs were dependability, MOS-specific task performance, MOS knowledge, military bearing, performance under adverse conditions, and performance on common, general soldiering tasks, e.g., putting on a face mask. The construct weights for the 3 MOS were assigned by the officers under a replicated 3 X 3 Graeco Latin square design in which three weighting procedures were used under three different military scenarios (see Figure 1).

The three weighting procedures employed all involved direct judgments of the relative weight that each performance construct should receive in forming an overall composite score based on all six constructs. In procedure A, the officers were first asked to rank order the 6 constructs. They were then told to assign 100 points to the first ranked construct and to scale the other constructs accordingly. (This is a variant of the Kane method.) In procedure B, the officers were instructed to divide 100 points among the 6 constructs in a manner that reflected the relative weight that should be given the constructs in forming the composite performance measure. In procedure C, 15 pairs of the 6 factors were presented in a paired comparison protocol.[1] The officers' task was to divide 100 points between the two constructs being judged in any given pair.

[1] In this paired comparison protocol and others used in this research the order of the presentations of the pairs was governed by the optimization procedure worked out by Ross (1934).

The judgments were made in the context of three different scenarios (see Figure 2). The scenarios described respectively a peacetime condition, a period of heightened tensions, and a wartime setting in which hostilities had just broken out. The site (i.e., Europe) of the three scenarios was the same.

No. of Subjects	11B	63W	71L
2	Aa	Bb	Cc
1	Bc	Ca	Ab
1	Cb	Ac	Ba
	63W	**71L**	**11B**
2	Aa	Bb	Cc
2	Bc	Ca	Ab
2	Cb	Ac	Ba
	71L	**11B**	**63W**
2	Aa	Bb	Cc
2	Bc	Ca	Ab
2	Cb	Ac	Ba

Scaling methods: Maximum 100 points (A), Divide 100 points (B),
Paired comparison (C).

Military scenario: Wartime (a), period of heightened tensions (b),
peacetime (c),

Figure 1. Replicated GRAECO-Latin Square Design

PEACETIME SCENARIO

Europe is in the peacetime condition currently prevailing there. Your Corps' mission is to defend and maintain the host country's border should war break out. The potential enemy approximates a combined arms Army and has nuclear and chemical capability. Air parity does exist. The Corps has personnel and equipment sufficient to make it mission capable for training and evaluation. The training cycle includes periodic field exercises, command and maintenance inspections, ARTEP evaluations, and individual soldier training/SQT testing.

WARTIME SCENARIO

Hostilities have broken out in Europe and your Corps' combat units are engaged. Your Corps' mission is to defend, then reestablish, the host country's border. Pockets of enemy airborne/heliborne and guerilla elements are operating throughout the Corps sector area. Limited initial and reactive chemical strikes have been employed but nuclear strikes have not been initiated. Air parity does exist.

HEIGHTENED TENSIONS SCENARIO

Europe is in a period of heightened tensions. There is an increasing probability that hostilities will break out in the next several months. Your Corps' mission is to defend and maintain the host country's border should war break out. The potential enemy approximates a combined arms Army and has nuclear and chemical capability. Air parity does exist. The Corps' training and other preparatory activities have been substantially increased. Most combat and associated support units are participating in frequent field exercises. Most units are being actively resupplied.

Figure 2. Three Different Military Scenarios

After completing the construct weighting judgments called for by each weighting procedure/scenario/MOS combination, each officer completed an *evaluation form* in which s/he rated the weighting method (procedure plus scenario) s/he had just employed. The evaluation form contained four 7-point scales which allowed the officers to rate the weighting methods on four dimensions:

1) acceptability to the Army,

2) ease of making the judgments called for by the method,

3) their confidence in the validity of the judgments made, and

4) the amount of agreement with other workshop participants that could be expected.

After all officers had completed their ratings, a short informal discussion period was held in which the opinions of the officers about the methods was solicited.

The research design permitted testing for the significance of differences in mean ratings on the four dimensions of the procedures and scenarios and for the significance of any procedure X scenario interactions. Neither the procedure mean differences nor the scenario mean differences were significant on the four scales taken separately or averaged into one combined index. However, significant ($p < .05$) procedure X scenario interactions were obtained for the acceptability to the Army, confidence in judgment validity, and the average composite index. Examination of the means obtained for the nine procedure/scenario combinations (see Table 1) revealed that procedure A (in which 100 points were assigned to the first ranked construct) had particularly low ratings when combined with the peacetime scenario, but had relatively high ratings when combined with the wartime and heightened tension scenarios.

Experiment 1: Mean Ratings across Four Dimensions of Nine Weighting Procedure/Scenario Combinations

(separate means based on ratings of 5 or 6 officers)

	Scenario		
Procedure	Peacetime	Heightened Tensions	Wartime
A. Maximum = 100 pts.	2.85	4.75	4.79
B. Divide 100 pts.	4.95	5.12	4.20
C. Paired Comparison	4.62	4.60	4.35

In the discussions following the administration of the construct weighting methods, the officers generally expressed preference for procedures A and C over procedure B. They felt that the time they spent worrying about whether the sum of their weights equalled 100 detracted from their ability to judge the relative importance of the weights. The officers also expressed a general preference for the heightened tensions and wartime scenarios over the peacetime scenario as the military setting for the judgments. They felt that the primary purpose of the Army was to prevent the outbreak of war and to carry out its missions successfully if there was a war.

Although the results of the statistical tests did not support the choice of one procedure or scenario over another, we felt that if a larger number of constructs were ultimately identified by Project A analyses of the criterion space, procedures B and C (divide 100 points and paired comparisons) could become fairly onerous. Procedure A, on the other hand, would be relatively easy to apply even if as many as nine or ten constructs had to be weighted. We, furthermore, felt that some of the NCO subject matter experts, who we were planning to use along with officers in actually collecting the construct weights for the Project A MOS, might find procedures B and C difficult to apply.

Non-statistical considerations also were used in choosing between the heightened tensions and wartime scenarios. Primarily, we felt that a heightened tension scenario would evoke a more uniform frame of reference or judgmental setting across the many different kinds of SME providing the MOS construct weights than a wartime scenario would unless the wartime scenario was made quite specific. However, specificity in the scenario could produce unwanted dependency of the construct weights on particular elements in the scenario, which could detract from the validity of the weighted composite as an overall, general measure of MOS performance. Furthermore, in a parallel research effort, we had decided to use a heightened tension scenario to collect information concerning the relative utility to the Army of different levels of performance in different MOS. (Earlier research had indicated that judgments of soldiers' utility were impacted by scenario differences [Sadacca and Campbell, 1985]). As these utility judgments would later be used in conjunction with the construct weights in further development of the Army's enlisted personnel selection and classification system, it was felt that the same scenario should be used as the setting for both types of judgments.

Experiment 2 - Procedure and Results

The second weighting method field experiment was conducted in Winter, 1986, at Fort Bragg, NC, using two four-hour workshops. One workshop was attended by 15 officers, the other by 15 NCOs. The workshop participants were asked to weight 5 performance constructs for the Infantry MOS: demonstrating commitment to the Army, maintaining technical proficiency and knowledge, maintaining physical fitness and military bearing, performance under adverse conditions, and maintaining and servicing weapons and equipment. Each of the participants used 3 different weighting methods:

1) Rank order the 5 constructs, assign 100 points to the first ranked construct, and then scale the other constructs accordingly (procedure A in Experiment 1).

2) Based upon their scores on the separate constructs, rank order 25 infantrymen in order of their overall performance. (For each of the infantrymen, a different set of performance scores on the 5 constructs was given on 7-point scales that range from the lowest level of performance to the highest. See Figure 3.)

3) Based upon their scores on 2 constructs, rank order 10 sets of 13 infantrymen in order of their overall performance. (In each set, the performance scores on 2 constructs are given on the same 7-point scales used in the second method above. A set of 13 infantrymen is given for each of the 10 possible pairs of the 5 constructs. See Figure 4).

The second and third methods are variants of the conjoint approach to scaling in which instead of obtaining the relative importance of the performance constructs directly through judgments, the judges' weights for the performance constructs are inferred from the rank order given sets of hypothetical soldiers whose performance on the constructs has been systematically varied. Multiple regression weights are calculated from the interrelationships between the rank orders provided by the judges and the performance construct levels given in the performance descriptions. In the paired comparison method, these regression weights are then used to derive the construct weights using a ratio scaling procedure described by Torgerson (1958, pp. 105-112). This procedure results in a set of scale values or weights for the constructs whose geometric mean is equal to 1.0.

The judgments were made in the context of a world-wide increase in tensions (see Figure 5). The weighting methods were applied in counterbalanced order by the 15 participants in each workshop. After completing each method, the participants rated the methods on the four 7-point scales used in the first experiment -- acceptability to the Army, ease of making the judgments, confidence in the validity of the judgments, and expected agreement with other participants.

Table 2 presents the mean ratings given the three weighting methods by the 30 workshop participants along with the results of analysis of variance tests of the significance of the method mean differences. The ratings clearly favored the direct estimation method, while the full profile conjoint method, which involved rank ordering the descriptions of 25 hypothetical infantrymen, generally received the lowest ratings. A breakout of these ratings by type of judge indicated that both the officers and NCOs generally preferred the direct estimation method most and the conjoint full profile method least.

The methods were also compared on three other dimensions: judge reliability (intraclass correlation), correlation between mean weights assigned by the officers and NCOs, and the intercorrelations among the sets of mean weights obtained by the three methods for all participants. These statistics are shown in Table 3. In general, the conjoint paired comparison

Soldier ____

Rank Order ________

Overall Score ________

MOS: Infantryman (11B)

A. **DEMONSTRATING COMMITMENT TO THE ARMY**
Maintaining Army traditions, spirit and fellowship.

Shows lack of dedication to Army traditions and values.	Generally supports Army traditions and values.	Shows constant devotion to Army tradition and values.
1 2	3 4 5	6 7

B. **PHYSICAL FITNESS AND MILITARY APPEARANCE**
Maintaining military standards of physcial fitness; maintaining proper military appearance and standards of cleanliness and grooming.

Maintains self in poor physical condition. Fails to meet military standards for dress and personal hygiene.	Meets Army standards of physical fitness. Dresses neatly and meets Army standards of personal hygiene.	Exceeds Army standards and expectations set for physical fitness. Maintains excellent personal hygiene and proper appearance.
1 2	3 4 5	6 7

C. **MAINTAINING AND SERVICING WEAPONS AND EQUIPMENT**
Keeping weapons and equipment clean and serviced and prepared for the field.

Fails to perform or improperly performs checks and preventive maintenance on weapons and equipment.	Performs routine checks and preventive maintenance on weapons and equipment.	Always keeps assigned weapons and equipment in ready-for-inspection condition.
1 2	3 4 5	6 7

D. **TECHNICAL PROFICIENCY AND KNOWLEDGE**
Effectiveness in applying technical knowledge and proficiency in carrying out MOS tasks.

Does not display the knowledge/skill required to perform many job assignments and tasks.	Displays the knowledge/skill required to perform most job assignments and tasks properly, but may need help for harder tasks	Displays the knowledge/skill to perform all job assignments and tasks properly.
1 2	3 4 5	6 7

E. **PERFORMANCE UNDER ADVERSE CONDITIONS**
Continuing to execute appropriate soldier skills under combat conditions or under hardship, stressful or otherwise difficult circumstances.

Makes frequent mistakes in combat situations or otherwise stressful situations.	Makes mistakes infrequently in combat or otherwise stressful situations.	Almost never makes mistakes in combat or otherwise stressful situations.
1 2	3 4 5	6 7

Judge No.______

Figure 3. Example of Rank Order Score Sheets

Name________________________________ Sheet No. 01

OVERALL PERFORMANCE SCORE SHEET

	Performance Level			
Soldier No.	Demonstrating Commitment to the Army	Technical Proficiency and Knowledge	Rank Order	Overall Score
1	5	5		
2	1	4		
3	2	6		
4	4	7		
5	4	4		
6	6	5		
7	6	2		
8	3	2		
9	4	1		
10	5	6		
11	2	3		
12	3	3		
13	7	4		

Performance Scales:

DEMONSTRATING COMMITMENT TO THE ARMY

Maintaining Army traditions, spirit and fellowship.

<table>
<tr><td colspan="2">Shows lack of dedication to Army traditions and values.</td><td colspan="3">Generally supports Army traditions and values.</td><td colspan="2">Shows constant devotion to Army tradition and values.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

TECHNICAL PROFICIENCY AND KNOWLEDGE

Effectiveness in applying technical knowledge and proficiency in carrying out MOS tasks.

<table>
<tr><td colspan="2">Does not display the knowledge/skill required to perform many job assignments and tasks.</td><td colspan="3">Displays the knowledge/skill required to perform most job assignments and tasks properly, but may need help for harder tasks</td><td colspan="2">Displays the knowledge/skill to perform all job assignments and tasks properly.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

Figure 4. Example of an Overall Performance Score Sheet

The world is in a period of heightened tensions. There is an increasing probability that hostilities will break out in Europe, Asia, and Caribbean, Latin American, and africa. The Army's mission is to support U.S. treaty obligations and to help defend the borders of allied and friendly nations. Some of the potential enemies have nuclear and chemical capability. Air parity does exist between allied forces and potential hostile nations. U.S. Army training and other preparatory activities have been substantially increased. Most combat and associated support units are participating in frequent field exercises. Most units are being actively resupplied.

Figure 5. World-Wide Increase in Tensions Scenario.

Table 2

Experiment 2: Mean Ratings of Weighting Methods

(n = 15 officers, 15 NCOs)

Weighting Method	Acceptability	Ease	Validity	Agreement	Average Rating
Direct estimation	4.30	5.13	5.80	4.77	5.00
Conjoint paired comparison	4.23	4.13	5.17	4.50	4.51
Conjoint full profile	4.27	3.87	5.10	4.23	4.37
Significance	.020	.002	.048	ns	.04

Table 3

Experiment 2: Agreement Indexes for Weighting Methods

	One-Rater Reliability			Correlation	Intercorrelations	
Weighting Method	Officer	NCO	All	Off/NCO Means	Full Profile	Paired Comp
Direct estimation	.27	.24	.25	.81	.17	.93
Conjoint full profile	.23	.01	.11	.60		.15
Conjoint paired comparison	.54	.32	.42	.91		

method yielded the highest intraclass correlations for both the officers and NCOs while the conjoint full profile method had the lowest values. The correlation between the mean officer and NCO weights obtained from the conjoint paired comparisons method also was the highest (r =.91), while the conjoint full profile officer/NCO correlation was the lowest (r = .60). The mean weights obtained from the direct estimation and the conjoint paired comparison methods were highly correlated (r = .93) while the correlations of these weights with those obtained from the conjoint full profile method were quite low. On the basis of these results and the participant method evaluations described earlier, it was decided to drop the conjoint full profile method from further consideration.

Experiment 3 - Procedure and Results

The third weighting method field experiment was also conducted in Winter, 1986, at Ft. Bragg, NC, using two four-hour workshops. One workshop was attended by 7 officers, the other by 8 NCOs. The workshop participants were asked to weight 7 performance constructs for the Infantry MOS. The 7 constructs included the 5 used in the second weighting method experiment plus 2 additional ones--avoiding serious disciplinary problems and providing peer leadership and support. Each of the participants used 3 different weighting methods in the following order:

1) Based on their scores on 2 constructs, rank order 21 sets of 13 infantrymen in order of their overall performance. (This is the same basic conjoint paired comparison procedure used in the second experiment. In this case, however, in addition to rank ordering the 13 infantrymen, the judges assigned overall performance scores to the soldiers by first assigning the top ranked soldier a score of 100 and then giving the remaining soldiers scores that reflected their relative overall performance.)

2) Rank order the 7 constructs, assign 100 points to the first ranked construct, and then scale the other constructs accordingly (the direct estimation procedure used in Experiments 1 and 2).

3) Indicate why the performance factors were ranked and weighted as they were in method 2 above. (These reasons are passed around to the other workshop participants. Also passed around are the average and range of the weights given each performance factor by the workshop participants in method 2.) After considering this feedback information, reassign weights to the performance factors using method 2 above. Repeat another round of feedback by indicating why the 7 performance factors were ranked as they were the second time and examining the new set of reasons given by the other participants and the new average and range of the weights. Then rank order and scale the performance factors for the third and last time. (In this modified Delphi technique, the judges were also told not to discuss the weights they assigned with the other workshop participants and that they could, if they wished, disregard the feedback when reassigning the weights).

The judgments were made in the same context of a world-wide increase in tensions that was used in Experiment 2 (Figure 5). After completing each method, the participants rated the methods on the same four 7-point scales used in the first and second experiments.

Table 4 presents the mean ratings given the 3 weighting methods by the 15 workshop participants along with the results of analysis of variance tests of the significance of the method mean differences. The ratings for the direct estimation and modified Delphi methods were generally higher than those given the conjoint paired comparison method. A breakout of these ratings by type of judge indicated that both the officers and NCOs generally preferred the conjoint method least while giving a slight edge to the Delphi over the direct estimation method given with no feedback.

Table 4

Experiment 3: Mean Ratings of Weighting Methods

(n = 7 Officers, 8 NCOs)

Weighting Method	Acceptability	Ease	Validity	Agreement	Average Rating
Conjoint paired comparison	3.43	4.20	4.60	3.86	4.02
Direct estimation	4.21	5.27	5.80	4.57	4.95
Modified Delphi	4.46	5.43	5.93	4.62	5.09
Significance	ns	.049	.010	ns	.002

It is interesting to note that the mean ratings given the direct estimation method in Experiments 2 and 3 (see Tables 2 and 4) were generally quite similar, while the conjoint paired comparison method generally received lower ratings in Experiment 3 than in Experiment 2, although only the mean acceptability ratings for this conjoint method were significantly different across the two experiments (4.23 vs. 3.43). The generally lower ratings received by the conjoint paired comparison method was not unexpected since there were 21 sets of 13 soldiers to be rank ordered in Experiment 3 while there were only 10 sets in Experiment 2. Also, the instructions for the conjoint procedure in Experiment 3 called for assigning overall performance scores to the 13 infantrymen in addition to rank ordering them.

The weighting methods used in Experiment 3 were also compared on the three dimensions used in Experiment 2 to compare the weighting methods: judge reliability (intraclass correlation), correlation between mean weights assigned by the officers and NCOs, and the intercorrelations among the sets of mean weights obtained by the three methods. Two of the methods, the conjoint paired comparison and Delphi methods, allowed two sets of construct weights to be derived from the judgments made by the workshop participants. For the conjoint paired comparison method weights could be derived only using the rank orders provided by the judges as the dependent variables when computing the regression weights. Construct weights could also be derived from the overall performance scores assigned the sets of 13 infantrymen. Similarly, for the modified Delphi method weights could be obtained from the participants' judgments after the first round of feedback or after the second and final round of feedback. One-rater reliabilities were therefore calculated for five different procedures of obtaining weights from the judgments provided by the workshop participants. These reliabilities, along with the correlations of the mean weights of the officer and NCO participants, are shown in Table 5. The correlations obtained between the five sets of mean weights are shown in Table 6. Also shown in Table 6 are the intercorrelations across weights of the five common constructs used in Experiments 2 and 3 for all the methods used in the two experiments.

Several inferences can be made from the data presented in Tables 5 and 6. First, there is no evidence that the one-rater reliabilities or the correlations obtained from the officers and the NCOs are improved substantially by adding the requirement to provide overall performance scores in addition to ranks in the conjoint paired comparison method. Nor are these agreement indexes improved by adding the requirement to obtain one or two rounds of Delphi feedback to the direct estimation method. Moreover, the correlations between weights obtained through the two basic methods (conjoint paired comparisons-ranking and direct estimation) and the weights obtained through their respective extensions (conjoint paired comparison-scores and Delphi-rounds 1 and 2) ranged from .96 to .99.

There was some evidence, however, that the correlations of the weights derived using the conjoint paired comparison overall performance scores with the weights devised using the other methods were higher than the corresponding correlations of the weights derived from the conjoint method just using the rank orders. For example, the correlation between the set of seven weights derived from the conjoint method using performance scores and those derived from the second Delphi feedback round was .80, while the corresponding correlation for the weights derived using the conjoint method rank orders was .64. These higher correlations were offset by the lower one-rater reliabilities found for the weights derived from the conjoint scores than for the weights derived from the conjoint rank orders (see Table 5).

Two other considerations, one practical, the other theoretical, led us to decide not to require that the judges assign overall performance scores in addition to rank ordering the sets of soldiers in any future application of the conjoint paired comparison method. From a practical point of view, the requirement to assign performance scores added about two minutes on the average to the amount of time it takes to complete the judgment for one set of 13 hypothetical soldiers. In Experiment 2, the workshop participants completed the 10 paired comparison sets in about 30 minutes on the average or 3 minutes per set. In Experiment 3, the workshop participants completed the 21 sets in an average of about 5 minutes per set. The heavier judgment demands on the participants may in part have led to the somewhat lower evaluations that the conjoint paired comparison method received in Experiment

Table 5

Experiment 3: Agreement Indexes for Weighting Methods

	One-Rater Reliability			Correlation Off/NCO
Weighting Method	Officer	NCO	All	Means
Conjoint PC-ranking	.43	.27	.35	.84
Conjoint PC-scores	.32	.20	.27	.87
Direct estimation	.28	.20	.25	.84
Delphi-round 1	.26	.18	.22	.75
Delphi-round 2	.32	.18	.24	.77

Table 6

Experiments 2 and 3: Intercorrelations of Mean Weights Obtained From the Weighting Methods Used in Both Experiments

Weighting Method	No. of Construct	Conjoint PC Ranking	Conjoint PC Scores	Direct Est.	Delphi Round 1	Delphi Round 2	Direct Est.	Conjoint Full Profile
Conjoint PC-ranking	7	-						
Conjoint PC-scores	7	.96	-					
Direct Estimation	7	.73	.86	-				
Delphi-round 1	7	.65	.80	.96	-			
Delphi-round 2	7	.64	.80	.99	.97	-		
Direct est. (Exp. 2)	5	.82	.91	.96	.93	.93	-	
Conj (full prof.-Exp. 2)	5	.12	.19	.36	.44	.44	.17	-
Conj. (paired Comp. Exp. 2)	5	.97	.98	.87	.87	.81	.93	.15

3 (see Tables 2 and 4). At any rate, the increased time and effort required to assign performance scores as well as ranks did not seem to be warranted psychometrically, especially if a large number of performance constructs emerged from the analysis of the criterion data. For example, with 8 constructs an additional 1-1/2 hours of judgment time might be needed to make sure most workshop participants completed the 28 sets of soldiers to be judged.

From a theoretical viewpoint, assigning overall performance scores to the sets of soldiers has a problem which assigning ranks does not have. The problem has to do with the assumption one makes about the soldiers' scores on the constructs that are not being immediately compared in the paired comparison protocol. The overall performance scores assigned the set of soldiers for the pair of constructs being judged might be different, if one assumes that these other construct scores are all high, than if one assumes that these scores are low, average, or mixed. The rank orders, on the other hand, should not be influenced by this assumption provided that all soldiers are assumed to have the same pattern of scores on the other constructs.

Similar considerations led us to decide not to use the modified Delphi method in addition to the direct estimation method. One might have expected greater agreement among the participants after receiving feedback, but the one-rater reliabilities obtained did not reflect this. Nor were the correlations of the officer and NCO mean weights and correlations with mean weights derived from other methods any higher for the Delphi method than for the direct estimation method. Although the Delphi method did receive slightly higher average ratings than the direct estimation method, the additional time (over 20 minutes per round on the average) taken to provide the feedback did not seem to be warranted in view of the other statistical results.

The choice between the direct estimation method and the conjoint paired comparison-ranking method was not an easy one. On the one hand, the direct estimation method generally received higher evaluation ratings in both Experiment 2 and 3 and would obviously take less time to administer than the conjoint method. (With the 5 constructs in Experiment 2, it took the workshop participants about 7 minutes on the average to complete the direct estimation method compared to 30 minutes for the conjoint method.) On the other hand, the officer and NCO one-rater reliabilities obtained for the conjoint method were higher in both experiments. Both the direct estimation and paired comparison methods had correlations between the officer and NCO mean weights above .80 in both experiments. The correlations between the mean weights obtained in Experiment 2 with those obtained in Experiment 3 were very high for both methods (.96 for the direct estimation and .97 for the conjoint method) indicating that the relative weights of the five common constructs were not unduly affected in either method by the increased number of constructs (7) used in the third experiment. In short, although one method might have some advantages over the other and vice versa, both appeared to be sound methods of obtaining performance construct weights. We therefore decided to use both methods to obtain the construct performance weights for the Project A MOS sample.

Performance Construct Weighting Workshops for Project A MOS

Method

The performance construct weighting workshops were originally scheduled to be completed in Spring 1986. However, as a result of various delays most of the workshops were conducted in June and July and some of the workshops have not been completed as of the date of this presentation. For each MOS, four 2-hour workshops were initially scheduled for collecting SME judgments concerning the relative importance of the construct weights. Two workshops, one for officers the other for NCOs, were scheduled at each of two Army posts. One of these posts housed the proponent school for the MOS while the other housed field units having officers and NCOs with expert knowledge of the MOS. The locations, dates, and numbers of officers and NCOs who attended the workshops are given in Table 7.

At each workshop, the participants were first given general instructions which covered the background, and purpose of the workshop and descriptions of the performance constructs and the two methods (direct estimation and conjoint paired comparison-ranking) that would be used to obtain weights for the constructs. Assumptions to be used in making the judgments were also given. These included the world-wide heightened tensions scenario (Figure 5) and the narrowing of the judgment basis to the performance of first-tour soldiers and the specific constructs to be weighted. The general instructions also gave definitions of the five performance constructs that were given to all MOS:

1) Maintaining personal discipline;
2) Military bearing/appearance and physical fitness;
3) Exercise of leadership, effort and self-development;
4) Task proficiency: MOS specific technical skills; and
5) Task proficiency: General soldiering skills.

An example[2] of the general and specific instructions for the two weighting methods (direct estimation and conjoint paired comparison-ranking) are given in Appendix A, along with the forms on which the workshop participants recorded their judgments.

After reading the general instructions and having any questions raised by the participants' answered by the workshop leader, the participants proceeded to read the specific instructions for the first method (direct estimation). After they completed this method and handed in their estimated weights, they proceeded to read the instructions for the conjoint method and perform the required judgments. The workshop leader, a member of HumRRO's Project A research staff, was available at all times to answer any questions that might arise.

Although the same 5 performance constructs were used for all MOS, their order of presentation in the instructions and on the judgment recordation forms was randomized across MOS. The order of the hypothetical soldiers listed on the 10 conjoint paired comparison sheets was the same, however, for all MOS. This order was determined randomly. Instead of presenting 13 hypothetical soldiers to be rank ordered we presented 15 soldiers in order to

[2] The instructions and forms were made MOS-specific, in the sense that the name of the MOS for which weights were being obtained was featured in each set.

Table 7

Schedule of Performance Construct Weighting Workshops

MOS	Location	Date(s)	Officers	NCO	Total
11B	USAREUR	16-27 Jun 86	10	6	16
	Ft. Benning	20 Aug 86			
12B	USAREUR	16-27 Jun 86	17	4	21
	Ft. Belvoir	10 Jul 86	9	6	15
	Ft. Belvoir	30 Jul 86	3	0	3
13B	USAREUR	16-27 Jun 86	0	6	6
	Ft. Sill	17 Sep 86			
16S	USAREUR	16-27 Jun 86	11	6	17
	Ft. Bliss	13 Aug 86			
19E/K	Ft. Knox	10 Jun 86	10	6	16
	Ft. Hood	21 Jul 86	7	1	8
27E	USAREUR	16-27 Jun 86	0	6	6
	Redstone Arsenal	24 Jun 86	8	5	13
31C	Ft. Gordon	6 Jun 86	12	6	18
	USAREUR	16-27 Jun 86	13	6	19
51B	Ft. Lewis	8 Jul 86	4	6	10
	Ft. Belvoir	10 Jul 86	7	4	11
	Ft. Belvoir	30 Jul 86	5	2	7
54E	USAREUR	16-27 Jun 86	9	8	17
	Ft. McClellan	19 Jun 86	12	6	18
55B	Redstone Arsenal	24 Jun 86	8	5	13
	Ft. Lewis	8 Jul 86	4	4	8
63B	USAREUR	16-27 Jun 86	7	2	9
	Aberdeen Proving Gd.	11 Jul 86	11	6	17
64C	Ft. Lewis	7 Jul 86	10	6	16
	Ft. Dix/Eustis	To be scheduled			
67N	Ft. Rucker	3 Jun 86	10	6	16
	Ft. Hood	22 Jul 86	12	1	13
71L	Ft. Ben Harrison	17 Jun 86	11	7	18
	Ft. Lewis	7 Jul 86	13	6	19
76W	USAREUR	16-27 Jun 86	0	6	6
76Y	Ft. Lee	16 Jul 86	10	6	16
	Ft. Hood	22 Jul 86	8	5	13
91A	USAREUR	16-27 Jun 86	12	6	18
	Ft. Sam Houston	To be Scheduled-Oct 86			
94B	USAREUR	16-27 Jun 86	4	6	10
	Ft. Lee	16 Jul 86	8	6	14
95B	Ft. McClellan	19 Jun 86	12	6	18
	Ft. Hood	21 Jul 86	12	6	18

increase by a small amount the number of cases upon which the regression equations would be computed (see Overall Performance Score Sheets, Appendix A). The scores of the 15 soldiers on any two constructs being evaluated were set so that their means were equal to 4.0 (the midpoint of the performance scales), their variances were equal, and the correlation between the construct scores was zero.

Preliminary Results

As of the date that this paper was prepared, we had processed and conducted preliminary analyses on the direct estimation and conjoint paired comparison data from 10 officer and 10 NCO workshops. The participants in these workshops were subject matter experts for the 10 MOS listed in Table 8. A total of 164 officers and NCOs participated in the 20 workshops. Examination of the 10 separate regression equations calculated from the conjoint paired comparison protocols of each participant revealed that 22 of the participants (6 officers and 16 NCOs) had regression equations in which one or more constructs had a positive weight. This would mean that the higher the participant rank ordered the 15 hypothetical soldiers in the given set, the lower was the soldiers' scores on that construct. As the ratio scaling method employed (see Torgerson, op. cit.) required both weights to have the same sign, we could either eliminate the conjoint protocols of the 22 participants or try to adjust their effective weights so that the constructs could be scaled. We decided to try adjusting the weights of those participants (n=13) who only had one equation with a positive weight. The adjustment was made by assuming the ratio of the negative weight to the positive weight in the equation was 99 to 1. This adjustment, in effect, assumes that the participants intended to give the positively weighted construct very little weight in comparison to the other construct.

Two sets of 1-rater and n-rater reliabilities were calculated for the conjoint paired comparison method in each MOS, one including the participants for whom the weighting ratio adjustments were made and the other excluding these cases. In general, both the 1-rater and n-rater reliabilities were lowered when the adjusted cases were included in the calculations. Therefore, these cases were dropped in further analyses of the conjoint method data.

Table 8 gives for the 10 MOS the 1-rater reliabilities (intraclass correlations) obtained for the officers, NCOs, and all judges combined for both the direct estimation and conjoint methods.

In general, these reliabilities were somewhat disappointing, especially those for the NCOs. For the direct estimation method, the NCO 1-rater reliabilities were .10 or below for 6 of the 10 MOS. For both weighting methods the NCO reliabilities were lower on the average than those of the officers, a result consistent with the results of the earlier field weighting experiments. If this trend continues as more workshop data is processed and analyzed, we may wish to question whether we should exclude the NCO data from the calculation of the final weights or at least weight these data separately. The n-rater reliabilities given in Table 9 for all workshop participants and for the officers above bear on this issue. Both sets of reliabilities averaged about .80 despite the increased number of raters involved in the calculation for the reliabilities for the combined groups.

Table 8

Estimated 1-Rater Reliabilities by MOS and Weighting Method
(Preliminary Data)

	Direct Estimation						Conjoint Paired Comparison					
	Officer		NCO		All		Officer		NCO		All	
MOS	No.	Rel.	No.	Rel.	No.	Rel.	No.	Rel.	No.	Rel.	No.	Rel.
Infantryman (11B)	10	.52	6	.01	16	.33	7	.53	5	.01	12	.37
MANPADS Crewman (16S)	11	.38	6	.01	17	.15	11	.32	3	.62	14	.31
Armor Crewman (19E)	10	.34	6	.40	16	.32	9	.64	5	.22	14	.46
TOW/Dragon Repairer (27E)	8	.10	5	.00	13	.06	7	.15	4	.32	11	.20
Radio Teletype Operator (31C)	12	.23	6	.08	18	.21	12	.07	5	.28	17	.10
NBC Specialist (54E)	12	.32	6	.40	18	.13	12	.20	4	.17	16	.09
Ammunition Specialist (55B)	8	.59	6	.39	14	.35	8	.27	4	.38	12	.20
Utility Helicopter Repairer (67N)	10	.51	6	.10	16	.34	10	.44	5	.13	15	.25
Administrative Specialist (71L)	11	.33	7	.04	18	.20	11	.25	4	.18	15	.15
Military Police (95B)	12	.41	6	.22	18	.35	11	.53	5	.11	16	.33

Table 9

Estimated n-Rater Reliabilities by MOS and Weighting Method
(Preliminary Data)

	Direct Estimation						Conjoint Paired Comparison					
	Officer		NCO		All		Officer		NCO		All	
MOS	No.	Rel.	No.	Rel.	No.	Rel.	No.	Rel.	No.	Rel.	No.	Rel.
Infantryman (11B)	10	.91	6	.07	16	.89	7	.89	5	.02	12	.88
MANPADS Crewman (16S)	11	.87	6	.07	17	.74	11	.84	3	.83	14	.87
Armor Crewman (19E)	10	.83	6	.80	16	.88	9	.94	5	.59	14	.92
TOW/Dragon Repairer (27E)	8	.47	5	.00	13	.44	7	.56	4	.66	11	.74
Radio Teletype Operator (31C)	12	.79	6	.34	18	.83	12	.47	5	.66	17	.65
NBC Specialist (54E)	12	.85	6	.80	18	.73	12	.75	4	.44	16	.62
Ammunition Specialist (55B)	8	.92	6	.79	14	.88	8	.75	4	.71	12	.75
Utility Helicopter Repairer (67N)	10	.91	6	.39	16	.89	10	.89	5	.42	15	.83
Administrative Specialist (71L)	11	.84	7	.22	18	.82	11	.79	4	.47	15	.72
Military Police (95B)	12	.89	6	.62	18	.90	11	.93	5	.37	16	.89

The correlations between the mean weights of the officers and NCOs also bear on this issue. These correlations are given in Table 10. For both weighting methods, the average across the 10 MOS of the correlations between the officer and NCO weights was less than .50 and was even negative in the case of 2 MOS. Considering the low correlations between the officer and NCO mean weights it is not surprising that combining their data did not result in higher n-rater reliabilities than just using the officer data.

Table 10 also presents the correlations between the mean weights derived from the two methods for the officers, NCOs and all cases combined. As might be expected from their respective reliabilities, the correlations between the mean weights derived from the direct estimation method and those derived from the conjoint method were generally higher for the officers than for the NCOs. More than half of the correlations for the officers and combined groups were above .90, indicating that the two methods were yielding similar sets of relative weights. These correlations, as well as the n-rater reliabilities, might be expected to increase when the data for these 10 MOS from the other weighting workshops are combined with the present data.

Table 11 presents the mean weights obtained from the direct estimation method weights for the five constructs for the officers and NCOs by MOS. Table 12 gives the comparable means derived from the conjoint paired comparison method. Separate analyses of variance were run on the weights derived from the two methods to test the significance of mean construct differences. Overall analyses using a repeated measures paradigm and separate analyses of variance by construct were run for the two methods. The separate analyses by construct were run to help identify the sources of the significant within subject interaction terms that were obtained in the overall analyses for the two methods. The probabilities of the F-tests run in these analyses are given in Tables 13 and 14.

The means in Tables 11 and 12 and the analysis of variance significance levels given in Tables 13 and 14 point to several important trends in these preliminary data. First, the mean weights assigned to the separate constructs by the workshop participants varied significantly across MOS for the most part. The overall construct 2nd MOS interaction term was highly significant for both the direct estimation and conjoint analyses, and the separate construct analyses indicated significant MOS mean differences in all but two cases (maintaining personal discipline and general soldiering skills for the conjoint method). For example, under both weighting methods, the MOS-specific technical skills construct was weighted considerably greater in the Ammunition Specialist MOS (55B), than it was in the Infantryman MOS (11B); while the leadership construct received higher weights for Infantryman than it did for Ammunition Specialists. Compared to other MOS, the military bearing/appearance and physical fitness construct was considered relatively unimportant for Tank Crewman (19E), while the maintaining personal discipline construct received relatively low weights on the average for Infantrymen.

Another important preliminary finding concerns the relative importance of the constructs across the MOS. In general, the MOS-specific technical skills construct received the highest weight and the military bearing/appearance and physical fitness construct received the lowest weight across the MOS, officers and NCOs, and weighting methods. In all 10 MOS the military bearing/appearance and physical fitness construct was weighted lowest on

Table 10

Selected Correlation Coefficients of Mean Weights by MOS
(Preliminary data)

	Direct Estimation	Conjoint Paired Comparison	Dir. Est./Conj. PC[1]		
MOS	Officer/NCO	Officer/NCO	Officer	NCO	All
Infantryman (11B)	.75	.95	.93	.74	.89
MANPADS Crewman (16S)	- .11	.53	.95	.57	.93
Armor Crewman (19E)	.61	.84	.90	.76	.92
TOW/Dragon Repairer (27E)	.58	.78	.90	.36	.73
Radio Teletype Operator (31C)	.93	.94	.70	.64	.69
NBC Specialist (54E)	- .27	- .25	.91	.63	.99
Ammunition Specialist (55B)	.31	.52	.95	.89	.88
Utility Helicopter Repairer (67N)	.73	.48	.93	1.00	.98
Administrative Specialist (71L)	.49	- .55	.93	-.60	.90
Military Police (95B)	.92	.32	.94	.55	.96

[1] The method correlations are based only on the data from the workshop participants who completed both methods successfully.

Table 11

Mean Direct Estimation Weights for Officers, NCOs, and Combined Groups by MOS

(Preliminary Data)

	MOS										
Construct	11B	16S	19E	27E	31C	54E	55B	67N	71L	95B	Total
Officers											
Maintaining personal discipline	50.5	76.8	58.5	84.8	77.3	71.7	61.3	82.5	83.2	84.2	73.4
Military bearing/appearance	41.5	57.1	46.5	64.8	69.5	60.0	38.8	54.0	49.5	70.8	55.9
Exercise of leadership, effort	88.5	90.5	91.0	76.3	72.3	81.5	70.0	90.0	68.6	86.7	81.8
MOS-specific technical skills	78.5	92.3	84.0	89.0	95.0	79.6	95.0	88.5	88.0	86.8	89.1
General soldiering skills	76.0	80.0	72.5	81.9	85.9	94.2	91.9	63.5	78.1	97.5	80.8
NCOs											
Maintaining personal discipline	79.2	94.2	67.5	78.0	80.8	90.0	80.0	82.5	84.3	75.0	81.3
Military bearing/appearance	60.8	91.7	45.0	67.0	70.8	86.7	87.5	72.5	79.3	70.0	73.3
Exercise of leadership, effort	83.3	87.5	65.0	79.0	67.9	57.5	65.8	80.0	68.6	81.3	73.4
MOS-specific technical skills	82.5	92.5	97.5	76.0	98.3	83.3	99.2	98.3	91.3	87.5	90.9
General soldiering skills	74.2	73.3	91.7	88.0	82.1	90.0	85.8	75.5	75.7	100.0	83.4
Combined											
Maintaining personal discipline	61.3	82.9	61.9	82.2	78.5	77.8	69.3	82.5	83.6	81.1	76.3
Military bearing/appearance	48.8	69.3	45.9	65.6	70.0	68.9	59.6	60.9	61.1	70.6	62.3
Exercise of leadership, effort	86.6	89.4	81.3	77.3	70.7	73.5	68.2	86.3	68.6	84.9	78.7
MOS-specific technical skills	80.0	92.4	89.1	84.0	96.2	90.6	96.8	92.2	89.3	87.1	89.8
General soldiering skills	75.3	77.6	79.7	84.2	84.6	83.1	89.3	68.0	77.2	98.3	81.7

Table 12

Mean Conjoint Paired Comparison Weights for Officers, NCOs, and Combined Groups by MOS

(Preliminary Data)

	MOS										
Construct	11B	16S	19E	27E	31C	54E	55B	67N	71L	95B	Total
Officers											
Maintaining personal discipline	.70	1.00	.78	1.31	.96	.98	1.00	1.43	1.15	1.10	1.05
Military bearing/appearance	.64	.60	.52	.70	.80	.63	.65	.60	.68	.60	.64
Exercise of leadership, effort	2.14	1.49	1.23	1.28	1.37	1.15	1.20	1.35	1.15	1.05	1.31
MOS-specific technical skills	1.31	1.65	1.94	1.35	1.52	1.40	1.35	1.22	1.29	1.03	1.40
General soldiering skills	1.23	0.95	1.22	.90	1.19	1.42	1.22	.88	1.10	1.91	1.22
NCOs											
Maintaining personal discipline	1.02	1.16	1.39	1.28	.99	1.16	.62	1.00	.91	1.10	1.06
Military bearing/appearance	.88	.77	.48	.75	.78	1.49	1.14	.72	1.07	.68	.86
Exercise of leadership, effort	1.29	1.81	1.09	1.26	1.08	.71	.65	1.13	1.06	1.59	1.16
MOS-specific technical skills	1.02	.93	1.94	.97	1.29	1.59	3.34	1.85	.95	1.12	1.51
General soldiering skills	1.05	.76	1.13	.96	1.02	.85	1.53	.93	1.05	1.12	1.05
Combined											
Maintaining personal discipline	.83	1.04	1.00	1.30	.97	1.02	.87	1.29	1.08	1.10	1.05
Military bearing/appearance	.74	.64	.50	.72	.79	.84	.82	.64	.78	.62	.71
Exercise of leadership, effort	1.79	1.56	1.18	1.27	1.29	1.04	1.01	1.27	1.13	1.22	1.26
MOS-specific technical skills	1.08	1.50	1.94	1.21	1.45	1.46	2.02	1.43	1.20	1.06	1.44
General soldiering skills	1.15	.91	1.19	.92	1.14	1.26	1.32	.90	1.09	1.66	1.16

Table 13

Analyses of Variance F-test Probabilities for the Direct Estimation Weights
(Preliminary Data)

			Separate Construct ANOVA				
Source	d.f.	Overall ANOVA	Personal Discipline	Military Bearing	Leadership Effort	MOS Tech. Skills	General Soldiering Skills
Between Subjects							
MOS	9	.0152	.0074	.0004	.0128	.0364	.0046
Group	1	.0098	.0170	.0001	.0184	.5193	.3507
MOS x Group	9	.6411	.1337	.0021	.5794	.3341	.6457
Error	143						
	162						
Within Subjects							
Constructs	4	.0001					
Constructs x MOS	36	.0001					
Constructs x Groups	4	.0001					
Construct X MOS x Groups	36	.0029					
Error	572						
	652						

Table 14

Analyses of Variance F-test Probabilities for the Conjoint Weights
(Preliminary Data)

			Separate Construct ANOVA				
Source	d.f.	Overall ANOVA	Personal Discipline	Military Bearing	Leadership Effort	MOS Tech. Skills	General Soldiering Skills
Between Subjects							
MOS	9	.0800	.2782	.0001	.0140	.0007	.0907
Group	1	.8159	.8004	.0001	.0964	.4311	.1130
MOS x Group	9	.0129	.1173	.0012	.1346	.0065	.4108
Error	122 141						
Within Subjects							
Constructs	4	.0001					
Constructs x MOS	36	.0001					
Constructs x Groups	4	.0284					
Construct X MOS x Groups	36	.0010					
Error	488 568						

the average under both weighting methods. In 6 of the 10 MOS, the MOS-specific technical skills construct was weighted highest on the average under both methods. The relative weights given the other three constructs were not as consistent across MOS and weighting method. In general, for the direct estimation method, the general soldiering skills and maintaining personal discipline constructs received the second and third highest average weights respectively. For the conjoint method, the second highest average weights were generally given to the exercise of leadership, effort and self-development construct, while the general soldiering skill^c construct received the third highest weights.

The differences in results by weighting method may be due, in part, to the greater number of NCOs represented in the direct estimation results. (A disproportionately larger number of NCOs than officers was dropped from the analyses of the conjoint weights, see page 19, blue nos.). Compared to the officers, the NCOs tended to give lower weights, on the average, to the leadership construct and higher weights to the maintaining personal discipline construct. NCOs also gave higher weights than officers on the average to the construct, military bearing/appearance and physical fitness. It may be that NCOs, in dealing with first-tour soldiers on a more daily basis, are emphasing different performance attributes than officers.

The data upon which these preliminary findings are based will be supplemented shortly by data from the remaining workshops. At that time, in addition to repeating the analyses described above for all sample MOS, analyses will be run comparing the relative weights obtained by weighting method and type of post (field vs. proponent school). We will also determine the impact of the obtained weight differentials on the validation of the ASVAB and the Project A trial battery and the use of these measures in selection and differential prediction. General conclusions concerning the weighting of Army performance construct scores to arrive at a measure of overall MOS performance will be drawn.

APPENDIX

General Instructions

JUDGING THE IMPORTANCE OF PERFORMANCE FACTORS IN ARRIVING AT TOTAL SCORES

Background

A number of different kinds of performance factors are being considered by Project A to assess the effectiveness of first-tour enlisted personnel. These various performance factors must be combined into one overall measure of MOS performance. This overall measure should be the best that can be obtained given the available component performance factors. The overall measure will be used as the performance measure against which the ASVAB and other predictor performance factors will be validated. To obtain the best overall measure for each MOS in our sample, Project A staff will be asking knowledgeable officers and NCOs to assign weights to the various performance factors in a manner that reflects the relative importance that the component performance factors should have in forming an overall measure for the MOS.

Today we would like to get your judgments about the relative weights that the factors should receive in deriving an overall performance measure for first-tour Infantryman (11B). The performance factors are:

> Task Proficiency: MOS specific technical skills--This performance factor represents the proficiency with which the soldier performs the tasks which are "central" to MOS 11B. The tasks represent the core of the job and they are the primary definers of the MOS. For example, the first tour Infantryman engages enemy target with hand grenades; installs and fires/recovers an M18A1 claymore mine; selects hasty firing positions in urban terrain; zeros an AN/PVS-4 to an M16A1 rifle; and uses weapons and other equipment in offensive and defensive combat operations.
>
> This performance factor does not include the individual's willingness to perform the task or the degree to which the individual can coordinate his efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so.

Task Proficiency: General soldiering skills--In addition to the core technical content specific to an MOS, individuals in every MOS are also responsible for being able to perform a variety of general soldiering tasks--for example, determines grid coordinates on military maps, puts on, wears and removes M17 series protective mask with hood, determines a magnetic azimuth using a compass, collects/reports information - SALUTE, recognizes and identifies friendly and threat aircraft. Performance on this factor represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldiering tasks, given a willingness to do so.

Exercise of Leadership, Effort, and Self Development--This performance factor reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient. While appropriate knowledges and skills are necessary for successful performance, this factor is only meant to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers.

Maintaining Personal Discipline--This performance factor reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self control, demonstrates integrity in day to day behavior, and does not create disciplinary problems. People who rank high on this factor show a commitment to high standards of personal conduct.

Military Bearing/Appearance and Physical Fitness--This performance factor represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

Please assume that a total score will be derived for each soldier from the separate scores obtained from each of these factors. These total scores will be our best estimate of the overall effectiveness of the troops whose performance will be measured. We need the assistance of experienced Army personnel in determining how much weight should be given each factor in arriving at the total effectiveness scores.

Purpose

The purpose of this workshop is to obtain the weights to be assigned each of the performance factors. Two methods of assigning weights will be used. The methods differ in the kinds of judgments you will be required to make:

Method A: You will be asked to rank order the performance factors and then assign weights to them, assuming that the top ranked factor has a weight of 100.

Method B: You will be given performance profiles on 10 sets of 15 soldiers each and asked to rank order them. (The profiles will give the scores of the soldiers on two of the five performance factors at a time).

Assumptions for Both Methods

(1) The type of soldiers for whom performance factor weights are being derived is first tour Infantryman (11B).

(2) As the weights you assign may be a function of the particular context in which the soldiers' performance is being evaluated, please assume the following military situation prevails:

> The world is in a period of heightened tensions. There is an increasing probability that hostilities will break out in Europe, Asia, the Caribbean, Latin America, and Africa. The Army's mission is to support U.S. treaty obligations and to help defend the borders of allied and friendly nations. Some of the potential enemies have nuclear and chemical capability. Air parity does exist between allied forces and potential hostile nations. U.S. Army training and other preparatory activities have been substantially increased. Most combat and associated support units are participating in frequent field exercises. Most units are being actively resupplied.

(3) Performance factor scores are available only on the factors given. Although there may be other factors that comprise overall performance, no scores are available for them at this time.

Materials for Method A

DIRECTIONS FOR METHOD A

Under this weighting method, the procedure for assigning weights to the performance factors is as follows:

1. Rank order the set of performance factors to be weighted by assigning a "1" to the most important, a "2" to the next most important, etc. Please refer to the "PERFORMANCE FACTORS FOR MOS 11B" handout for a complete description of the 5 performance factors.

2. After you have recorded the rank orders on the weighting sheet, assign 100 points to the factor you ranked as most important. Then ask yourself, "If I'm assigning 100 points to this performance factor, how many points should I assign to the next most important one." If, for example, you think that the second most important one should receive half the weight of the first, assign it 50 points. Continue assigning points in this manner until all the factors have been weighted.

3. In assigning the points, please keep in mind that the points represent how many times more (or less) important one performance factor is than another. For example, if you assign 30 points to one factor and 5 points to another, that means that you believe that the 30-point factor should receive 6 times the weight in the total score as the 5-point factor.

4. If you feel that two or more factors should be weighted equally you may assign them equal weights. For example, if you feel that the factors ranked first and second are really tied in importance, then you can assign them both 100 points.

5. If you believe that a particular performance factor should not be used at all in arriving at the total score, you should assign it zero points.

6. When you are finished assigning points to all performance factors please make sure that they are in the "right" ratio to one another. That is, the points assigned to all factors are in correct proportion to one another.

Thank you for your cooperation.

Name ______________ Workshop __________

MOS 11B Performance Factor Weighting Sheet

Performance Factor*	Rank Order	Weight
1. Task proficiency--MOS specific technical skills.	___	___
2. Task proficiency--general soldiering skills.	___	___
3. Exercise of leadership, effort, and self development.	___	___
4. Maintaining personal discipline.	___	___
5. Military bearing/appearance and physical fitness.	___	___

* Please refer to the "PERFORMANCE FACTORS FOR MOS 11B" handout for a complete description of the 5 performance factors.

PERFORMANCE FACTORS FOR MOS 11B

1) Task Proficiency: MOS specific technical skills

This performance factor represents the proficiency with which the soldier performs the tasks which are "central" to MOS 11B. The tasks represent the core of the job and they are the primary definers of the MOS. For example, the first tour Infantryman engages enemy target with hand grenades; installs and fires/recovers an M18A1 claymore mine; selects hasty firing positions in urban terrain; zeros an AN/PVS-4 to an M16A1 rifle; and uses weapons and other equipment in offensive and defensive combat operations.

This performance factor does not include the individual's willingness to perform the task or the degree to which the individual can coordinate his efforts with others. It refers to how well the individual can execute the core technical tasks the job requires, given a willingness to do so.

2) Task Proficiency: General soldiering skills

In addition to the core technical content specific to an MOS, individuals in every MOS are also responsible for being able to perform a variety of general soldiering tasks--for example, determines grid coordinates on military maps, puts on, wears and removes M17 series protective mask with hood, determines a magnetic azimuth using a compass, collects/reports information - SALUTE, recognizes and identifies friendly and threat aircraft. Performance on this factor represents overall proficiency on these general soldiering tasks. Again, it refers to how well the individual can execute general soldiering tasks, given a willingness to do so.

3) Exercise of Leadership, Effort, and Self Development

This performance factor reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers. That is, can the individual be counted on to carry out assigned tasks, even under adverse conditions, to exercise good judgment, and to be generally dependable and proficient. While appropriate knowledges and skills are necessary for successful performance, this factor is only meant to reflect the individual's willingness to do the job required and to be cooperative and supportive with other soldiers.

4) Maintaining Personal Discipline

This performance factor reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self control, demonstrates integrity in day-to-day behavior, and does not create disciplinary problems. People who rank high on this factor show a commitment to high standards of personal conduct.

5) Military Bearing/Appearance and Physical Fitness

This performance factor represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.

Materials for Method B

DIRECTIONS FOR METHOD B

Under this method, judgments of the overall performance scores for 10 sets of Infantrymen will be obtained. Each set will contain 15 Infantrymen. The performance scores of each of the 15 first tour Infantrymen have been recorded on 2 performance factor scales. (A different pair of performance factor scales are provided for each of the 10 sets). For each scale there is a description of high, medium and low levels of performance. Each of the 15 soldiers is rated on a 7-point scale that ranges from the lowest level of performance to the highest. Please refer to the "PERFORMANCE FACTORS FOR MOS 11B" handout for a complete description of the 5 performance factors. Also, please review the assumptions given on page 4 of the General Instructions.

Specific Instructions

1. Rank the 15 Infantrymen in the first set in order of their overall performance. Give the "best" soldier a rank of "1", the second best soldier a rank of "2" and so on. Make comparisons between the soldiers on the basis of their overall performance as Infantrymen; do not consider how they might be used in other capacities.

2. When you are finished, please go over the rank order carefully making sure that, in your judgment, the ranks reflect the relative overall performance of the soldiers. Feel free to change any ranks.

3. When satisfied with your rank ordering, proceed to the next set of 15 Infantrymen.

Thank you for your cooperation.

Name______________________________ Sheet No. 1

MOS 11B OVERALL PERFORMANCE SCORE SHEET

Performance Level

Soldier No.	Task Proficiency--MOS Specific Technical Skills	Task Proficiency--General Soldiering Skills	Rank Order
1	6	2	
2	5	5	
3	2	6	
4	5	3	
5	2	3	
6	6	5	
7	4	7	
8	1	4	
9	5	6	
10	7	4	
11	3	5	
12	3	3	
13	4	4	
14	4	1	
15	3	2	

Performance Scales:

TASK PROFICIENCY--MOS SPECIFIC TECHNICAL SKILLS

<table>
<tr><td colspan="2">Does not display the knowledge/skill required to perform many core technical skills.</td><td colspan="3">Displays the knowledge/skill required to perform most core technical tasks properly, but may need help for harder tasks.</td><td colspan="2">Displays the knowledge/skill to perform all core technical tasks properly.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

TASK PROFICIENCY--GENERAL SOLDIERING SKILLS

<table>
<tr><td colspan="2">Does not display the knowledge/skill required to perform many general soldiering tasks.</td><td colspan="3">Displays the knowledge/skill required to perform most general soldiering tasks, but may need help for harder tasks.</td><td colspan="2">Displays the knowledge/skill to perform all general soldiering skills.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

Name____________________ Sheet No. 2

MOS 11B OVERALL PERFORMANCE SCORE SHEET

Performance Level

Soldier No.	Military Bearing/ Appearance and Physical Fitness	Exercise of Leadership, Effort and Self Development	Rank Order
1	6	2	____
2	5	5	____
3	2	6	____
4	5	3	____
5	2	3	____
6	6	5	____
7	4	7	____
8	1	4	____
9	5	6	____
10	7	4	____
11	3	5	____
12	3	3	____
13	4	4	____
14	4	1	____
15	3	2	____

Performance Scales:

MILITARY BEARING/APPEARANCE AND PHYSICAL FITNESS

Maintains self in poor physical condition. Fails to meet military standards for dress or personal hygiene.	Meets Army standards of physical fitness. Dresses neatly and meets Army standards of personal hygiene.	Exceeds Army standards and expectations set for physical fitness. Maintains excellent personal hygiene and proper appearance.
1 2	3 4 5	6 7

EXERCISE OF LEADERSHIP, EFFORT AND SELF DEVELOPMENT

Fails to take charge when leadership is required in unit. Provides little or no assistance to other unit members. Seldom exerts effort in accomplishing many job assignments and tasks. Gives up easily under adverse conditions.	Performs satisfactorily in leadership situations where what is expected is well known. When asked, gives help and support to fellow soldiers. Usually exerts effort to perform most job assignments and tasks.	Takes charge when necessary to lead unit; leads the squad to outstanding performance. Does everything possible to assist other soldiers. Always exerts considerable effort in performing all job assignments and tasks.
1 2	3 4 5	6 7

Name ______________________ Sheet No. 3

MOS 11B OVERALL PERFORMANCE SCORE SHEET

	Performance Level		
Soldier No.	Maintaining Personal Discipline	Task Proficiency--MOS Specific Technical Skills	Rank Order
1	6	2	
2	5	5	
3	2	6	
4	5	3	
5	2	3	
6	6	5	
7	4	7	
8	1	4	
9	5	6	
10	7	4	
11	3	5	
12	3	3	
13	4	4	
14	4	1	
15	3	2	

Performance Scales:

MAINTAINING PERSONAL DISCIPLINE

Occasionally shows disrespect towards superiors; often fails to follow Army/unit rules, regulations or orders. Creates disciplinary problems.	Rarely exhibits disrespectful behavior towards superiors. Almost always follows Army/unit rules, regulations or orders.	Always treats superiors with respect. Maintains high level of personal integrity. Obeys orders quickly and with enthusiasm.
1 2	3 4 5	6 7

TASK PROFICIENCY--MOS SPECIFIC TECHNICAL SKILLS

Does not display the knowledge/skill required to perform many core technical skills.	Displays the knowledge/skill required to perform most core technical tasks properly, but may need help for harder tasks.	Displays the knowledge/skill to perform all core technical tasks properly.
1 2	3 4 5	6 7

Name________________________ Sheet No. 4

MOS 11B OVERALL PERFORMANCE SCORE SHEET

Performance Level

Soldier No.	Exercise of Leadership, Effort, and Self Development	Task Proficiency--General Soldiering Skills	Rank Order
1	6	2	
2	5	5	
3	2	6	
4	5	3	
5	2	3	
6	6	5	
7	4	7	
8	1	4	
9	5	6	
10	7	4	
11	3	5	
12	3	3	
13	4	4	
14	4	1	
15	3	2	

Performance Scales:

EXERCISE OF LEADERSHIP, EFFORT AND SELF DEVELOPMENT

Fails to take charge when leadership is required in unit. Provides little or no assistance to other unit members. Seldom exerts effort in accomplishing many job assignments and tasks. Gives up easily under adverse conditions.	Performs satisfactorily in leadership situations where what is expected is well known. When asked, gives help and support to fellow soldiers. Usually exerts effort to perform most job assignments and tasks.	Takes charge when necessary to lead unit; leads the squad to outstanding performance. Does everything possible to assist other soldiers. Always exerts considerable effort in performing all job assignments and tasks.
1 2	3 4 5	6 7

TASK PROFICIENCY--GENERAL SOLDIERING SKILLS

Does not display the knowledge/skill required to perform many general soldiering tasks.	Displays the knowledge/skill required to perform most general soldiering tasks, but may need help for harder tasks.	Displays the knowledge/skill to perform all general soldiering skills.
1 2	3 4 5	6 7

Name ____________________ Sheet No. 5

MOS 11B OVERALL PERFORMANCE SCORE SHEET

	Performance Level		
Soldier No.	Maintaining Personal Discipline	Military Bearing/ Appearance and Physical Fitness	Rank Order
1	6	2	
2	5	5	
3	2	6	
4	5	3	
5	2	3	
6	6	5	
7	4	7	
8	1	4	
9	5	6	
10	7	4	
11	3	5	
12	3	3	
13	4	4	
14	4	1	
15	3	2	

Performance Scales:

MAINTAINING PERSONAL DISCIPLINE

Occasionally shows disrespect towards superiors; often fails to follow Army/unit rules, regulations or orders. Creates disciplinary problems.		Rarely exhibits disrespectful behavior towards superiors. Almost always follows Army/unit rules, regulations or orders.			Always treats superiors with respect. Maintains high level of personal integrity. Obeys orders quickly and with enthusiasm.	
1	2	3	4	5	6	7

MILITARY BEARING/APPEARANCE AND PHYSICAL FITNESS

Maintains self in poor physical condition. Fails to meet military standards for dress or personal hygiene.		Meets Army standards of physical fitness. Dresses neatly and meets Army standards of personal hygiene.			Exceeds Army standards and expectations set for physical fitness. Maintains excellent personal hygiene and proper appearance.	
1	2	3	4	5	6	7

Name ______________________ Sheet No. 6

MOS 11B OVERALL PERFORMANCE SCORE SHEET

	Performance Level		
Soldier No.	Task Proficiency--MOS Specific Technical Skills	Exercise of Leadership, Effort, and Self Development	Rank Order
1	6	2	____
2	5	5	____
3	2	6	____
4	5	3	____
5	2	3	____
6	6	5	____
7	4	7	____
8	1	4	____
9	5	6	____
10	7	4	____
11	3	5	____
12	3	3	____
13	4	4	____
14	4	1	____
15	3	2	____

Performance Scales:

TASK PROFICIENCY--MOS SPECIFIC TECHNICAL SKILLS

Does not display the knowledge/skill required to perform many core technical skills.	Displays the knowledge/skill required to perform most core technical tasks properly, but may need help for harder tasks.	Displays the knowledge/skill to perform all core technical tasks properly.
1 2	3 4 5	6 7

EXERCISE OF LEADERSHIP, EFFORT AND SELF DEVELOPMENT

Fails to take charge when leadership is required in unit. Provides little or no assistance to other unit members. Seldom exerts effort in accomplishing many job assignments and tasks. Gives up easily under adverse conditions.	Performs satisfactorily in leadership situations where what is expected is well known. When asked, gives help and support to fellow soldiers. Usually exerts effort to perform most job assignments and tasks.	Takes charge when necessary to lead unit; leads the squad to outstanding performance. Does everything possible to assist other soldiers. Always exerts considerable effort in performing all job assignments and tasks.
1 2	3 4 5	6 7

Name________________________________ Sheet No. 7

MOS 11B OVERALL PERFORMANCE SCORE SHEET

	Performance Level		
Soldier No.	Task Proficiency--General Soldiering Skills	Maintaining Personal Discipline	Rank Order
1	6	2	____
2	5	5	____
3	2	6	____
4	5	3	____
5	2	3	____
6	6	5	____
7	4	7	____
8	1	4	____
9	5	6	____
10	7	4	____
11	3	5	____
12	3	3	____
13	4	4	____
14	4	1	____
15	3	2	____

Performance Scales:

TASK PROFICIENCY--GENERAL SOLDIERING SKILLS

Does not display the knowledge/skill required to perform many general soldiering tasks.	Displays the knowledge/skill required to perform most general soldiering tasks, but may need help for harder tasks.	Displays the knowledge/skill to perform all general soldiering skills.
1 2	3 4 5	6 7

MAINTAINING PERSONAL DISCIPLINE

Occasionally shows disrespect towards superiors; often fails to follow Army/unit rules, regulations or orders. Creates disciplinary problems.	Rarely exhibits disrespectful behavior towards superiors. Almost always follows Army/unit rules, regulations or orders.	Always treats superiors with respect. Maintains high level of personal integrity. Obeys orders quickly and with enthusiasm.
1 2	3 4 5	6 7

Name ____________________ Sheet No. 8

MOS 11B OVERALL PERFORMANCE SCORE SHEET

	Performance Level		
Soldier No.	Military Bearing/ Appearance and Physical Fitness	Task Proficiency-- MOS Specific Technical Skills	Rank Order
1	6	2	
2	5	5	
3	2	6	
4	5	3	
5	2	3	
6	6	5	
7	4	7	
8	1	4	
9	5	6	
10	7	4	
11	3	5	
12	3	3	
13	4	4	
14	4	1	
15	3	2	

Performance Scales:

MILITARY BEARING/APPEARANCE AND PHYSICAL FITNESS

<table>
<tr><td colspan="2">Maintains self in poor physical condition. Fails to meet military standards for dress or personal hygiene.</td><td colspan="3">Meets Army standards of physical fitness. Dresses neatly and meets Army standards of personal hygiene.</td><td colspan="2">Exceeds Army standards and expectations set for physical fitness. Maintains excellent personal hygiene and proper appearance.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

TASK PROFICIENCY--MOS SPECIFIC TECHNICAL SKILLS

<table>
<tr><td colspan="2">Does not display the knowledge/skill required to perform many core technical skills.</td><td colspan="3">Displays the knowledge/skill required to perform most core technical tasks properly, but may need help for harder tasks.</td><td colspan="2">Displays the knowledge/skill to perform all core technical tasks properly.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

Name________________________________ Sheet No. 9

MOS 11B OVERALL PERFORMANCE SCORE SHEET

	Performance Level		
Soldier No.	Exercise of Leadership, Effort, and Self Development	Maintaining Personal Discipline	Rank Order
1	6	2	____
2	5		____
3	2	6	____
4	5	3	____
5	2	3	____
6	6	5	____
7	4	7	____
8	1	4	____
9	5	6	____
10	7	4	____
11	3	5	____
12	3	3	____
13	4	4	____
14	4	1	____
15	3	2	____

Performance Scales:

EXERCISE OF LEADERSHIP, EFFORT AND SELF DEVELOPMENT

Fails to take charge when leadership is required in unit. Provides little or no assistance to other unit members. Seldom exerts effort in accomplishing many job assignments and tasks. Gives up easily under adverse conditions.	Performs satisfactorily in leadership situations where what is expected is well known. When asked, gives help and support to fellow soldiers. Usually exerts effort to perform most job assignments and tasks.	Takes charge when necessary to lead unit; leads the squad to outstanding performance. Does everything possible to assist other soldiers. Always exerts considerable effort in performing all job assignments and tasks.
1 2	3 4 5	6 7

MAINTAINING PERSONAL DISCIPLINE

Occasionally shows disrespect towards superiors; often fails to follow Army/unit rules, regulations or orders. Creates disciplinary problems.	Rarely exhibits disrespectful behavior towards superiors. Almost always follows Army/unit rules, regulations or orders.	Always treats superiors with respect. Maintains high level of personal integrity. Obeys orders quickly and with enthusiasm.
1 2	3 4 5	6 7

Name ______________________________ Sheet No. 10

MOS 11B OVERALL PERFORMANCE SCORE SHEET

Performance Level

Soldier No.	Task Proficiency--General Soldiering Skills	Military Bearing/Appearance and Physical Fitness	Rank Order
1	6	2	____
2	5	5	____
3	2	6	____
4	5	3	____
5	2	3	____
6	6	5	____
7	4	7	____
8	1	4	____
9	5	6	____
10	7	4	____
11	3	5	____
12	3	3	____
13	4	4	____
14	4	1	____
15	3	2	____

Performance Scales:

TASK PROFICIENCY--GENERAL SOLDIERING SKILLS

<table>
<tr><td colspan="2">Does not display the knowledge/skill required to perform many general soldiering tasks.</td><td colspan="3">Displays the knowledge/skill required to perform most general soldiering tasks, but may need help for harder tasks.</td><td colspan="2">Displays the knowledge/skill to perform all general soldiering skills.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

MILITARY BEARING/APPEARANCE AND PHYSICAL FITNESS

<table>
<tr><td colspan="2">Maintains self in poor physical condition. Fails to meet military standards for dress or personal hygiene.</td><td colspan="3">Meets Army standards of physical fitness. Dresses neatly and meets Army standards of personal hygiene.</td><td colspan="2">Exceeds Army standards and expectations set for physical fitness. Maintains excellent personal hygiene and proper appearance.</td></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
</table>

References

Alpert, Mark I. Betak, John F., and Golden, Linda L. (1978). Data gathering issues in Conjoint measurement. Working Paper. Graduate School of Business, Stanford University.

Bernardin, H. John and Beatty, Richard W. (1984). Performance Appraisal: Assessing Human Behavior at Work. Kent Publishing Company: Belmont, California.

Green, Paul E. (1974) On the design of choice experiments involving multifactor alternatives. The Journal of Consumer Research, 1, 61-73.

Green, Paul E. and Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. Journal of Consumer Research, 5, 103-123.

Jain, Arun K., Acito, Franklin, Malhotra, Naresh, and Mahajan, Vijay (1978). A comparison of predictive validity of alternative methods for estimating parameters in preference models. Working Paper, School of Management, State University of New York, at Buffalo.

Johnson, Richard M. (1974). Trade-off analysis of consumer values, Journal of Marketing Research, 11, 121-127.

Johnson, Richard M. and VanDyk, Gerald J. (1975). A resistance analogy for efficiency of paired comparisons designs. Unpublished Paper. Market Facts, Inc., Chicago.

Kane, J.S. (1980). Performance distribution assessment: A new framework for conceiving and appraising job performance. Unpublished manuscript, Kane and Associates.

Montgomery, David B., Wittink, Dick, R. and Glaze, Thomas (1977). A predictive test of individual level concept evaluation and trade-off analysis. Research Paper No. 415. Graduate School of Business, Stanford University.

Oppedijk van Veen, Walle M. and Beazley, David (1977). An investigation of alternative methods of applying the trade-off model. Journal of Market Research Society, 19, 2-9.

Ross, R.T. (1934). Optimum orders for presentation of pairs in pair comparisons. Journal of Educational Psychology, 25, 375-382.

Schmidt, F.L. (1977). The measurement of job performance. Unpublished manuscript, U.S. Office of Personnel Management.

Torgerson, W.S. (1958). Theory and methods of scaling. New York: John Wiley & Sons, Inc.

A PATH ANALYTIC MODEL OF JOB PERFORMANCE RATINGS

Leonard A. White
U.S. Army Research Institute

Walter C. Borman and Leaetta M. Hough
Personnel Decisions Research Institute

R. Gene Hoffman
Human Resources Research Organization

Presented on symposium,
"Causal Models of Job Performance"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

This paper is based on data collected for a large Army research project titled, "Improving the Selection, Classification, and Utilization of Army Enlisted Personnel: Project A" (Eaton, Hanser, & Shields, 1980). The primary goal of this effort is to increase Army organizational effectiveness by improving the soldier-job match. This objective will be achieved by constructing a broad set of selection and classification measures (predictors) and job performance criteria and empirically investigating relationships between these predictor and criterion measures.

The performance measurement blueprints for Project A called for the development of (a) hands-on, task proficiency tests, (b) paper-and-pencil tests of job knowledge, and (c) supervisory and peer ratings of job performance. Rating measures were constructed to capture job-specific task performance and typical performance on a broader set of "Army-wide" effectiveness dimensions including personal discipline and motivation. Later in the project, the multiple measures of performance will be combined into a single composite or composites to measure an enlisted soldier's effectiveness on the job.

A long term objective of the present research is to use data from Project A to develop a model(s) of overall job performance ratings that incorporates a wide range of potential contributors to ratings. The notion is that by learning about factors that influence ratings and relationships between ratings and other kinds of criterion measures we will achieve a better understanding of the "meaning" of supervisory and peer ratings as measures of job performance.

The starting point for the present research was a study by Hunter (1983). In a meta-analysis of 14 studies, Hunter used causal analysis techniques to identify relationships among four variables relevant to work performance: cognitive ability, job knowledge, work samples, and ratings of job performance. The analysis showed that supervisor ratings were related to both job sample test scores and job knowledge required for effective performance, but these relationships were relatively low. In the model, cognitive ability influenced supervisory ratings of job performance by enabling workers to learn the facts, skills, and procedures required to do their jobs.

More recently, Borman, White, Gast, and Pulakos (1985) used data from five entry level Army jobs to examine relationships reported by Hunter (1983). The model of peer and supervisory rating reported by Borman et al., (1985) is presented in Figure 1. As can be seen, peer and supervisory ratings of overall job performance showed the strongest relationship to work sample performance and slightly weaker links to job knowledges required for effective performance. The multiple correlation for the prediction of ratings from job knowledge and work sample performance was .36 for supervisors and .28 for peer assessments. Thus, factors other than those included in the model would appear to account for a large portion of the variance in ratings.

The Present Research

This paper follows-up on the work of Borman et al. (1985) and Hunter (1983) with special attention to possible contributors to job performance ratings. Several factors have been proposed as having potential to influence ratings. Broadly speaking, these include characteristics of the rater and ratee, the context in which the appraised is conducted, and various rater and ratee interaction factors (Ilgen & Feldman, 1983; Landy & Farr, 1980).

We focus here on characteristics of ratees beyond work sample performance and job knowledge that may influence job performance ratings. Specifically, the models tested in this paper differed from Hunter (1983)

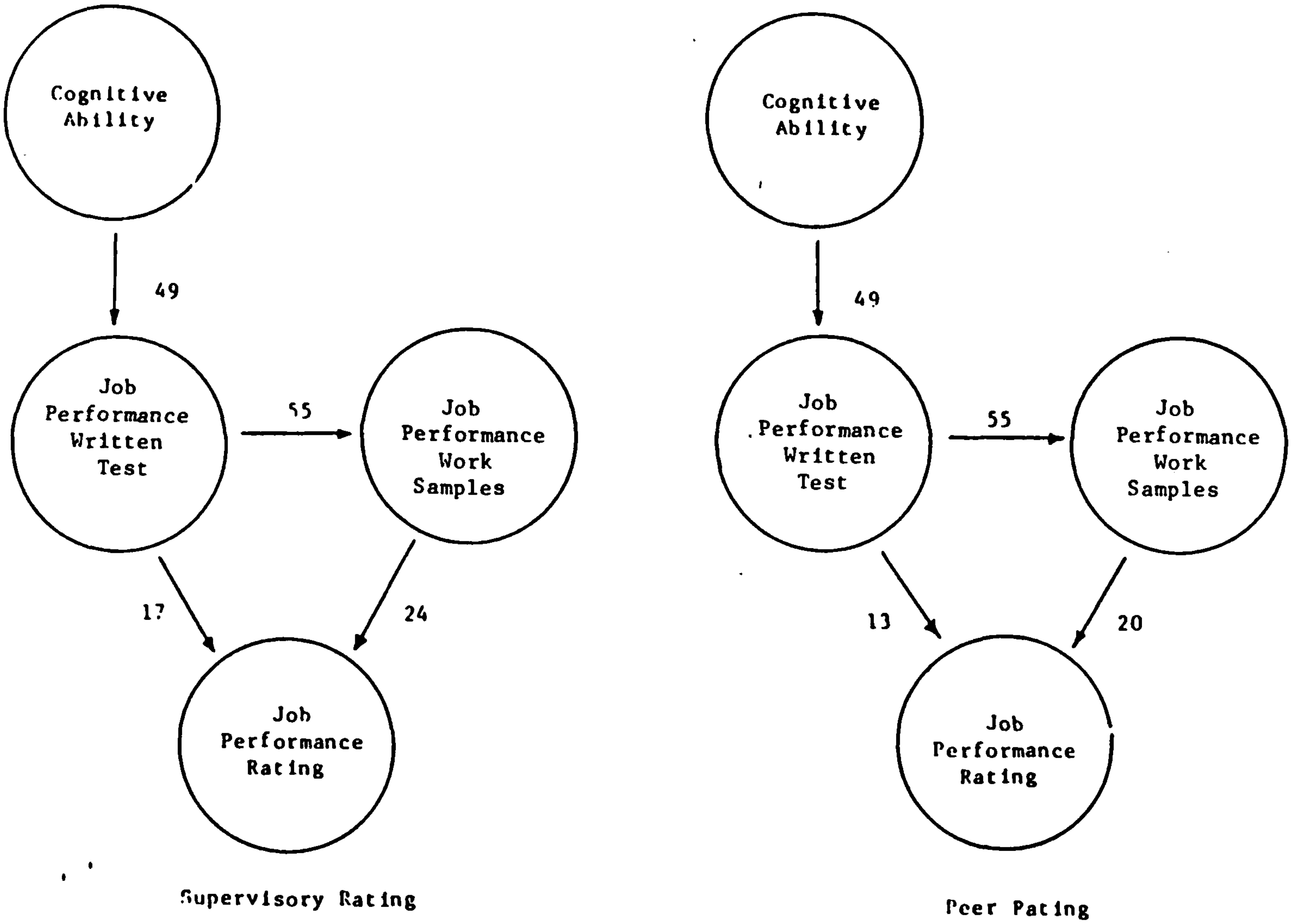


Figure 1. Model of overall job performance rating.

by including measures of ratee temperament and job experience. It was hypothesized that the pattern of relationships reported by Borman et al. (1985) would be obtained for the Army jobs examined in this research. In addition, several hypothesis were advanced regarding possible effects of ratee tempera-ment and job experience on overall job performance ratings.

Job Experience

Past research on relationships between job experience and performance has typically used supervisory ratings as the criterion measure. Based on a review of 425 investigations, Hunter and Hunter (1984) reported a mean correlation of .18 between months on the job (X = 61 months) and supervisory evaluations of performance. McDaniel (1985) found a higher correlation of .32 between experience in a manufacturing job (X = 43 months) and job performance ratings.

Schmidt, Hunter and Outerbridge (1986) report positive relationships between time in one's present job and performance capability as measured by work samples and job knowledge tests. The Schmidt et al. (1986) analysis was based on data from four Army jobs reported by Vineberg and Taylor (1972). They argue that the expected value of the correlation between job experience and performance is highest when experience is low (e.g., 0-3 years) due to greater inequality of experience among new job incumbents. For example, an employee who has worked for 3 months has only 8% as much experience as a person with 3 years of experience. A person with 16 years experience however, has 80% as much experience as someone who has worked 20 years. Evidence consistent with this notion of a curvilinear relationship between job experience and performance has been reported by McDaniel (1985).

The mean time on the job for the 1,530 soldiers in the present research was 19 months, with a relatively low standard deviation of six months and a range of 8-36 months. A positive impact of job experience on performance was expected however, albeit lower than the path coefficients based on the Vineberg and Taylor (1972) research where Army experience ranged from 1 month to 20 years. In the Vineberg and Taylor (1972) research, job experience had the strongest effect on job knowledge and indirect links to supervisory ratings of effectiveness (Schmidt et al., (1986). However, research by Borman et al., (1985) suggests that experience during the first term of enlistment may also contribute to overall effectiveness by enabling troops to learn how to "get along" and manage themselves on a daily basis so as to improve productivity; skills typically not captured by work samples or written performance tests. In this paper, we explore possible direct and indirect effects of job experience on overall job performance ratings.

There are of course reasons for expecting job experience to have direct effects on work sample performance and job knowledge. Research on human learning has long ago documented a positive relationship between practice and performance. Military research on this topic is somewhat equivocal (Spiker, Harper, & Hayes, 1985), however positive relationships between experience and performance have been reported. In the present research, self-report rating scales were constructed to assess the frequency and recency of a job incumbent's opportunities to perform important job tasks, including those for which work sample tests were prepared. For the jobs examined here it was hypothesized that practice on job-specific tasks would be positively related to scores on work samples and written performance tests for both high- and lower-aptitude recruits.

Temperament

Reviews summarizing the relationship between temperament variables and job performance have come to differing conclusions about the relationship between temperament and job performance. Guion and Gottier in their 1965 Personnel Psychology article conclude that there is little, if any, relationship between temperament and job performance. Ghiselli, on the other hand, concluded in his 1973 Personnel Psychology review article that temperament variables do indeed relate to job performance. In an earlier phase of the Project A research, Hough, Kamp, and Barge (1984) reviewed both published and unpublished literature and found that when temperament variables are categorized into constructs and job performance criteria are categorized into types of criteria, temperament constructs relate in a significant and important way to job performance criteria. They found, for example, that measures of achievement orientation correlate in the .30s with educational and training criteria and in the mid .20s with supervisory ratings of job performance. They also found that measures of emotional stability and dependability correlate in the .30s and .40s with adjustment criteria such as unfavorable military discharge, delinquency, and substance abuse. Further, dependability correlated .13 with job proficiency criteria. Scales designed to measure these temperament constructs were included in the present research. Examination of relationships between scores on the temperament scales and job performance will provide clues about the direct and indirect influence of these non-cognitive constructs on supervisory and peer ratings of overall job performance.

In summary, the purpose of the research was twofold: a) to evaluate relationships between ratings and other measures of job proficiency and performance, and b) to explore the direct and indirect influence of job experience, cognitive ability, and temperament factors on job performance ratings. In the analysis peer and supervisor raters were considered separately to expose possible effects of rating source on obtained relationships.

METHOD

Subjects

Participants in the research were 1530 first-term soldiers in three Army jobs; 600 military police (MOS 95B), 560 motor transport operators (MOS 64C), and 370 medical care specialists (MOS 91A).

Research Instruments

The first step in this research was to develop rating scales to measure performance on all relevant job factors and overall job performance in each of the three jobs. In addition, a job knowledge test and a hands-on, job sample test for 15 critical tasks was developed for each job.

Job performance rating scales. Critical incident workshops were conducted with Non-Commissioned Officers (NCO), first-line supervisors for each of the target jobs. The numbers of NCOs contributing critical incidents and the numbers of incidents generated were, 64C, 76 NCOs, 1147 incidents; 91A, 71 NCOs, 761 incidents, and 95B, 86 NCOs, 1183 incidents. Job-specific scales were developed for each job using a variate of the behaviorally anchored rating scale procedure (Smith & Kendall, 1963). These procedures resulted in seven to ten 7-point behavior summary scales for each of the five jobs (Toquam, McHenry, Corpe, Rose, Lammlein, Kemery, Borman, Mendel & Bosshardt, 1986). The rating scales focused on performance areas relevant to a specific job (e.g., patient care for the medical care specialist). In addition, a 7-point summary rating of overall job performance was included on the rating form. Scores on the overall job performance rating scale were averaged across peer raters and separately across supervisor raters. This aggregate performance measure provided the primary dependent variable for this research. For each job, Table 1 presents the mean number of peer and supervisor raters per soldier ratee.

Hands-on, task proficiency tests. For each of the jobs, 15 critical tasks representative of the entire task domain were the target for test development work. Task proficiency tests were prepared for each of the tasks (Campbell, Campbell, Rumsey & Edwards, 1986). Each task has several performance steps and each step was scored pass or fail. A proportion-passed score was derived for a soldier on each task and the proportions were averaged across tasks to yield an overall hands-on test score.

Written performance tests. Important knowledge areas for each job were carefully identified in job analysis work, and items intended to tap those knowledges were written with the help of subject matter experts (Campbell et al. 1986). For each soldier, the overall job knowledge test score was the percentage of correct answers on the test.

Job history questionnaire. A self-report instrument was developed to assess job opportunities to perform 28-30 MOS tasks during the six months prior to participation in this research. The tasks included all those for which job samples were prepared. Performance frequency for each task was reported on a categorical scale; not at all, 1-2 times, 3-5 times, 6-10 times, more than 10 times. The following scale was used to assess the most recent job opportunity to perform each task; during past month, 1-3 months ago, 4-6 months ago, more than 6 months ago, never. To provide a linear scale the frequency and recency categories were coded 1-5. A composite measure of task practice was computed for each ratee by summing the

TABLE 1

Summary of Supervisor and Peer Rater Assignments by Army Jobs

Variable	Motor Transport Operator	Medical Specialist	Military Police
Number of ratees	560	370	600
Mean number of supervisor raters/ratee	1.89	2.04	1.92
Mean number of peer raters/ratee	3.71	3.23	3.73

frequency and recency (reverse scored) data across tasks.

Cognitive ability. Prior to entrance into military service, ratees were administered the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is composed of 10 subtests and is used for selection and occupational classification. A composite measure of four ASVAB subtests known as the Armed Forces Qualification Test (AFQT) was used as a measure of cognitive ability.

Temperament scales. Ten temperament scales were developed by Hough, Kamp, and Barge to tap important elements of the constructs that had demonstrated criterion-related validity in previous studies (Hough, 1984). An inventory entitled "Assessment of Background and Life Experiences," (ABLE) was prepared and pilot-tested with a total of 470 people at three separate Forts. These data were used to revise the items and scales. Principal factor analysis of the revised scales scores with varimax rotation indicated that the 10 ABLE scales could be summarized by three factors with eigenvalues greater than one; Dependability, Achievement Orientation (Ascendency), and Emotional Stability (Hough & Ashworth, 1986). Factor scores were used as the measure of each ratee's score on the temperament constructs.

Briefly, the Ascendency factor incorporates measures of self-esteem, dominance, energy level, and wcrk orientation. High scores on this factor betoken self-confidence in one's abilities, a tendency to seek positions of leadership and a strong work ethic. The Emotional Stability factor relates to the degree of stability vs. reactivity of emotions. The emotionally stable person is generally calm, maintains an even mood, and maintains composure in a stressful situation. ABLE scales associated with the Dependability factor are indicative of conscientiousness, non-delinquency, support for rules and regulations, and respect for traditional values.

Procedure

Peer and supervisor raters were trained to use the behavior-based rating scales. With reference to the rater training literature (e.g., Bernardin & Pence, 1981; Pulakos, 1984), the training can be characterized as a combination of psychometric error and frame-of-reference program. The administrator described halo, restriction-of-range, and other rating errors in lay terms and provided guidance on how to reduce those errors. Also, the logic of the behavior-based scales was carefully explained, and raters were urged to study and then properly use the behavioral anchors to arrive at their ratings. After training, peer and supervisor raters in separate groups of 3-15 made their evaluations on the job performance scales. The first-term soldiers (ratees) were also administered the job knowledge and hands-on job sample tests. All participants were informed that the data gathered would be used only for research purposes.

Results

Table 2 presents the mean correlation between the work samples, written performance tests, ratee temperament, cognitive ability, job experience, and job performance ratings by peers and supervisors. The mean correlation across the three jobs was computed by weighting each correlation by its sample size. Meta-analysis techniques (Hunter, Schmidt, & Jackson, 1982) were used to determine the extent of non-artifactual variance around the average correlation. The correlations do vary somewhat across jobs, but much of the variance can be explained by sampling error. The range of correlation across jobs was almost always less than .10.

TABLE 2

Correlations Among Selecteᵈ Variables Across the Three Jobs

	1	2	3	4	5	6	7	8	9	10
1. Work Sample	.63									
2. Written Performance Test	.44*	.87								
3. Cognitive Ability	.20*	.35*	.93							
4. Achievement Orientation	.09*	.00	.11*	.80						
5. Dependability	.02	.14*	.02	-.04	.80					
6. Emotional Stability	.05	.10*	.06	-.03	-.01	.78				
7. Months in Service	.03	.01	-.01	.01	-.02	.08*	.90			
8. Task Practice	.20*	.07	.02	.15*	-.03	.05	-.01	.75		
9. Supervisor Rating	.21*	.17	.06	.19*	.14*	.08*	.14*	.06	.60	
10. Peer Rating	.15*	.10*	.01	.12*	.07	.06	.14*	.11*	.36*	.60

Note: (n=1530 ratees) Reliabilities are presented in the main diagonal.
* p < .01

Higher between-job variation was observed for correlations of dependability with job knowledge ($\underline{r}$ = .10 - .25) and dependability with peer ratings of job performance ($\underline{r}$ = .05 - .20).

As can be seen in Table 2, correlations between the cognitive and the non-cognitive predictors are relatively low. Time in service also showed low correlation with cognitive ability and the temperament measures. The highest correlations were observed among measures of cognitive ability, job knowledge and work sample performance. Further, task practice correlated significantly with work sample performance.

Overall job performance ratings by peers and supervisors evidenced low, positive correlations with achievement orientation, dependability, and measures of job knowledge and work sample performance. In general, relationships of job performance ratings with other variables were somewhat lower for peers, as compared to supervisors.

Correlations were also examined between the race of the ratee and peer and supervisory ratings. For the jobs examined here, correlations between overall job performance ratings and ratee race (arbitrarily coded White=1, Black=0) were all low ($.01 < \underline{r} < .07$) and not significantly different from zero across both rating sources.

Path Analyses

Path analysis was used to examine relationships among the variables presented in Table 2. Alternative models were tested using the LISREL VI program (Joreskog & Sorbom, 1985) which provides a maximum likelihood estimate of the model parameters. As an initial step in the data analysis, the research sample (n=1530) was divided into two independent random samples. One sample was used to test the proposed models and the second sample to verify the "accepted" model.

One concern here is that failure to correct correlation coefficients for measurement error can cause path coefficients to be biased (James, Mulaik, & Brett, 1982). To address possible effects of measurement error a procedure described by Kenny (1979) was used in the present research. This technique involves setting the path from each construct to its measured variable (λ) equal to the square root of the reliability of the measured variable. The amount of error variance (δ,ε) equals one minus the reliability.

The reliability estimates used in the analyses are presented in Table 2. Interrater agreement was used to estimate the reliability of job performance ratings. The reliability of the ABLE scales was based on a test-retest correlation over 2 to 8 weeks. Months in service was computed from the ratee's report of their service entry date. The reliability of this measure was assumed to be .90.

Supervisory Ratings Table 3 summarizes the models of supervisory rating tested for relationships in this research. Table 4 presents the degrees of freedom and results of the fit indicies for tests of the models described in Table 3. Rho (Bentler & Bonett, 1980) compares the fit of the estimated model relative to the null model. In general, values of rho over 0.90 indicate a good fit for the model (Harvey, Billings, & Nilan, 1985; James et al., 1982). The root mean square residual provides an overall index of the size of the difference between the fitted correlation matrix and the observed correlation matrix. Large residuals indicate that additional modifications may be called for to achieve a closer correspondence between the actual matrix and the fitted matrix.

The chi-square and associated probability value are also sensitive to departures of the model from the data. One concern here is that minor

TABLE 3

Model Descriptions

Model	Description of free causal paths

A	P → WS; JK → WS; T → WS;
	T → SR; ACH → SR; WS → SR; ES → SR; D → SR; JK → SR;
	AB → JK; P → JK; ACH → JK;
	ACH → P

B	P → WS; JK → WS;
	T → SR; ACH → SR; WS → SR; ES → SR; JK → SR;
	AB → JK; P → JK; D → JK; ES → JK;
	ACH → P

Structural Null	none

Note: T = time in service; WS = work sample; JK = job knowledge; SR = supervisory rating; P = job task practice; ACH = achievement orientation; D = dependability; AB = cognitive ability, ES = emotional stability.

discrepancies may be statistically significant (i.e., indicate poor fit) due to the dependency of the p-value on sample size. In the context of testing among alternative models, Joreskog & Sorbom (1979) advised that the ratio of the chi-square value to the degrees of freedom may be used as an indicator of when to stop fitting the model. When this ratio is large the model may be misspecified. When the ratio of the chi-square value to the degrees of freedom is small the model may be too restrictive and not generalize to other samples (Martin, Park, & Borman, 1986).

As summarized in Table 4, the model hypothesized by the authors (Model A) did not provide an adequate fit to the data. A close examination of Model A revealed three free structural parameters with T values that did not approach significance (i.e., time in service → work samples; job knowledge → rating; ascendency → job knowledge). Further, examination of LISREL modification indicies (MI) suggested a strong path from dependability to job knowledge (MI=14.57) and a weaker link from emotional stability to job knowledge (MI=6.37). The MI value represents the minimum expected decrease in the chi-square goodness of fit index if the constraint was relaxed.

The first step in the specification search was to free one at a time the paths from dependability and emotional stability to written performance tests. These modifications improved the goodness of fit as measured by the chi-square test. Next, several models were evaluated to determine if two of the three near zero paths could be omitted to improve parsimony without disturbing the overall fit of the model. The path from job knowledge to ratings was retained in the model because it has been supported in the literature (e.g., Hunter, 1983). Specifically, two nested models were evaluated to compare different orders of deleting parameters. The order

TABLE 4

Results of LISREL analysis

Model	df	Chi-square	rho	RMSR
		Supervisor Ratings		
A	25	49.58	.88	.039
B				
Sample 1	25	31.33*	.93	.031
Sample 2	25	26.49*	.94	.026
Structual null				
Sample 1	38	441.27		.122
Sample 2	38	427.73		.121
		Peer ratings		
B				
Sample 1	25	31.07*	.92	.030
Sample 2	25	31.32*	.92	.028
Structual null				
Sample 1	38	388.91		.113
Sample 2	38	399.99		.116

Note: rho compares the fit of the model relative to the null model; RMSR = root mean square residual.

* p > .05

of specification search had little effect on estimates of the remaining free parameters or the relative fit of the model. This resulted in Model B which provided a satisfactory fit to the data. Figure 2 shows the path coefficients estimated for Model B along with the correlations between cognitive ability and the temperament factors. The paths among the exogenous variables not shown in Figure 2 were restricted to zero.

A cross-validation of Model B yielded a reasonably good fit with $\chi^2(26, N=700) = 28.82$, $\underline{p} = .319$. Path coefficients for the model of supervisory ratings based on the entire sample are shown in Figure 3. The path estimating the direct impact of emotional stability on ratings did not even approach significance in the cross-validation sample and was therefore restricted to zero. In the model shown in Figure 3, the multiple correlation for the prediction of supervisory ratings was .52. With the exception of the path from written performance test scores to ratings the Ts for all nonzero path coefficients were significant (all $\underline{p} < .05$).

Peer ratings. Peers were more lenient in their ratings than supervisors in each of the Army jobs examined here. Results of F-tests indicated that these differences, however, did not reach statistical significance (all $\underline{p} > .05$).

Path analytic techniques were used to explore alternative models of peer ratings. Relatively little previous research has focused on relationships between measured performance capability, ratee temperament and overall job performance ratings by peers. One exception here is a study by Borman et al., (1985). They examined correlations between selected ratee traits (e.g., friendly), measured performance capability, and overall job performance ratings by both peers and supervisors. The pattern of correlation of job performance ratings with ratee social traits and

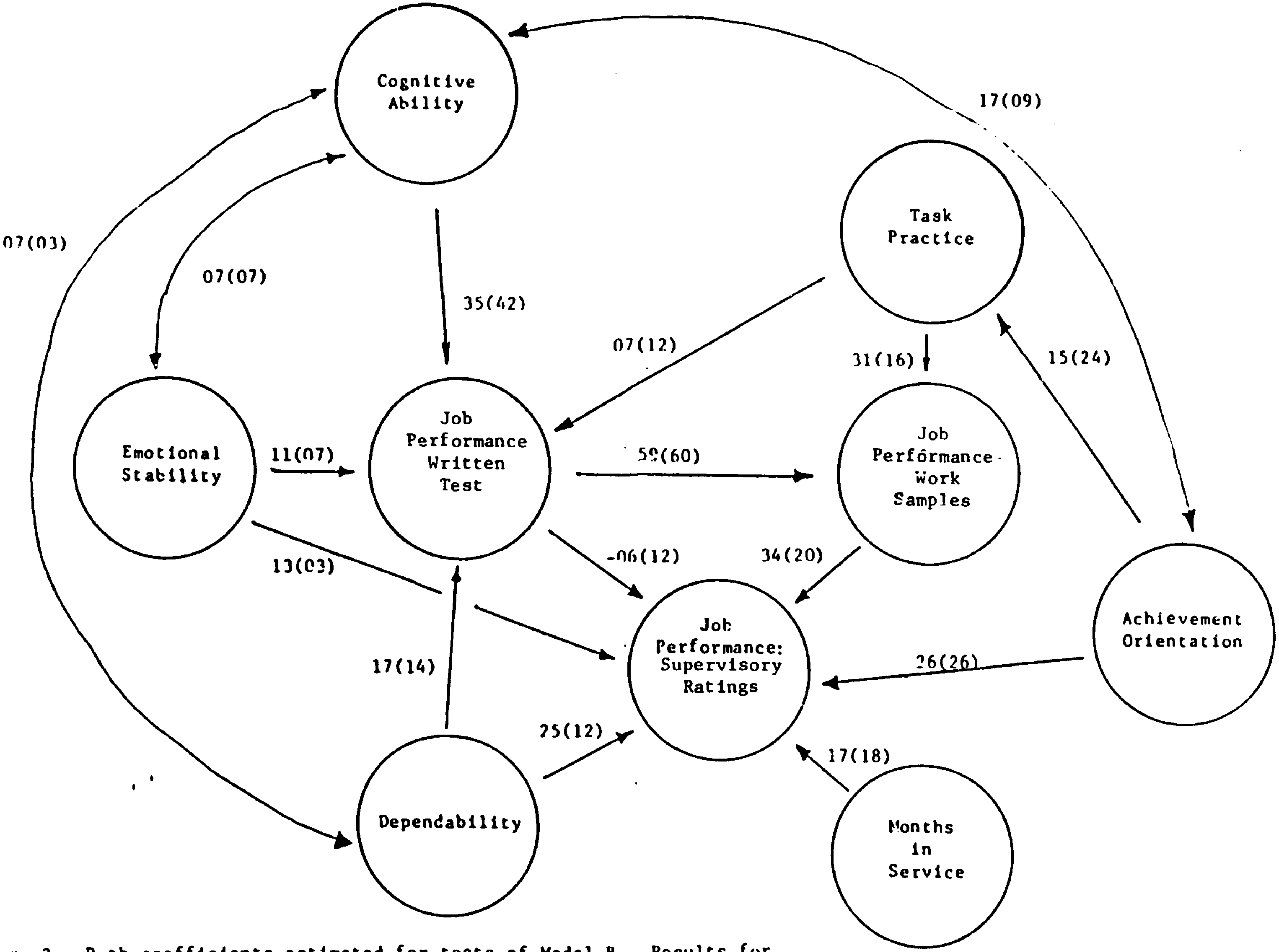


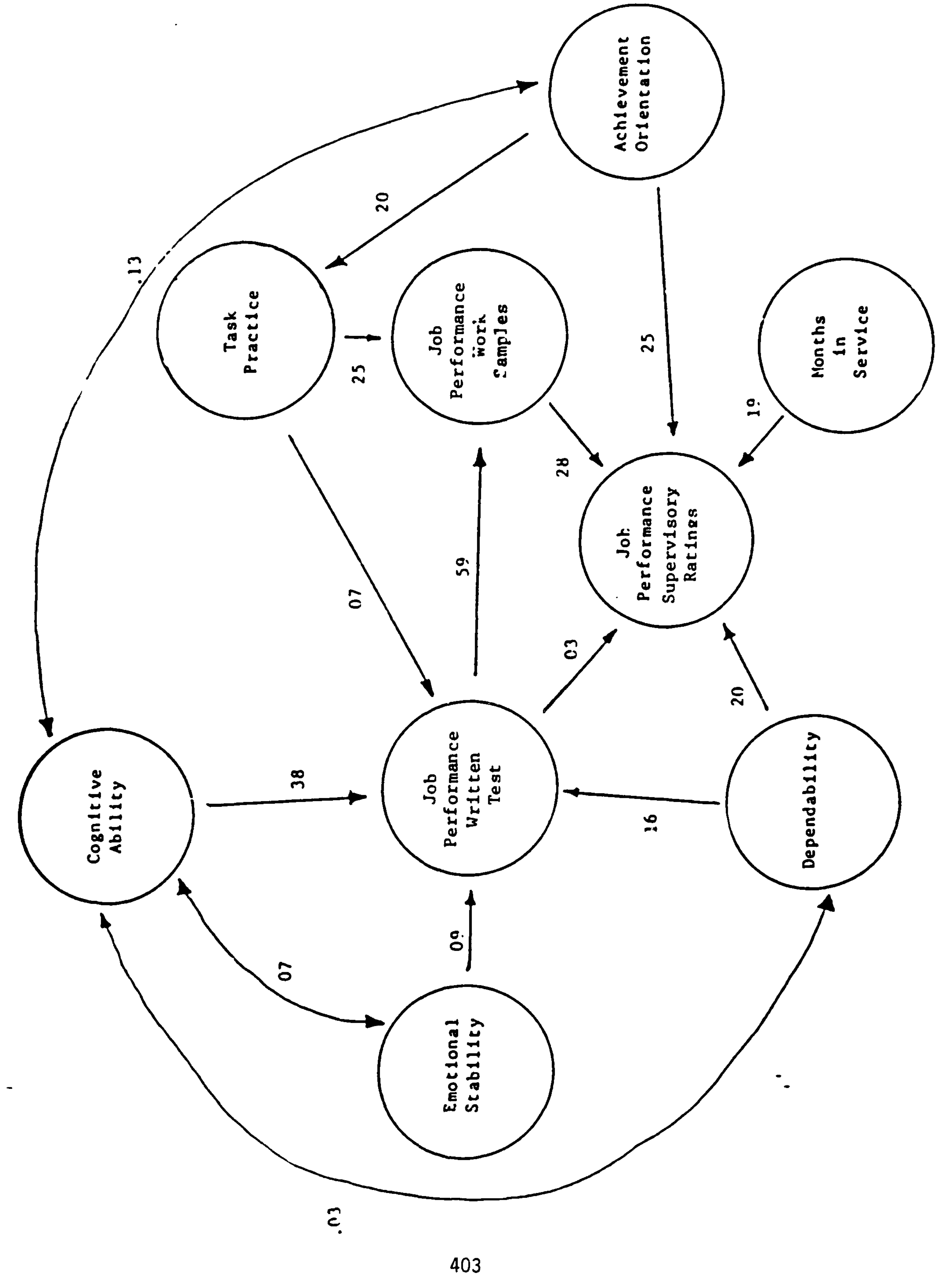
Figure 2. Path coefficients estimated for tests of Model B. Results for

demonstrated capability was very similar for peers and supervisors. Also of interest here, the multiple correlation for the prediction of overall job performance ratings from work samples and job knowledge was somewhat lower for peers, as compared to supervisors. A similar pattern of results was anticipated here.

To investigate this tentative hypothesis, the model of supervisory ratings shown in Figure 3 was applied to the peer rating data. Fit indicies and the degrees of freedom for tests of the model are presented in Table 4. Examination of these fit indicies and the fitted matrix residuals suggested that the model provided an acceptable fit. Path coefficients estimated for the model of peer ratings are shown in Figure 4. Once again, the paths from written performance tests → ratings and emotional stability → ratings were weak and not significantly different from zero. Removal of these two paths improved parsimony without disturbing the overall fit of the model. Parameter estimates for the "accepted" model based on the entire sample are presented in Figure 5. For the model presented in Figure 5, the multiple correlation for the prediction of peer ratings was .36. The T values for all nonzero paths exceeded the critical value of 2.0.

Discussion

The present research examined possible effects of job experience, cognitive ability, and measured performance capability on overall job performance ratings. As hypothesized, persons identified as achievement oriented, dependable, and committed to work received higher performance evaluations from their supervisors and to a lesser extent from their peers. Individual differences in dependability and ascendency impacted directly on job performance ratings as well as indirectly by contributing



Achievement Orientation
.13
20
Task Practice
Job Performance Work Samples
25
25
Months in Service
19
28
Job Performance Supervisory Ratings
07
59
03
20
Cognitive Ability
38
Job Performance Written Test
16
Dependability
07
09
Emotional Stability
.03

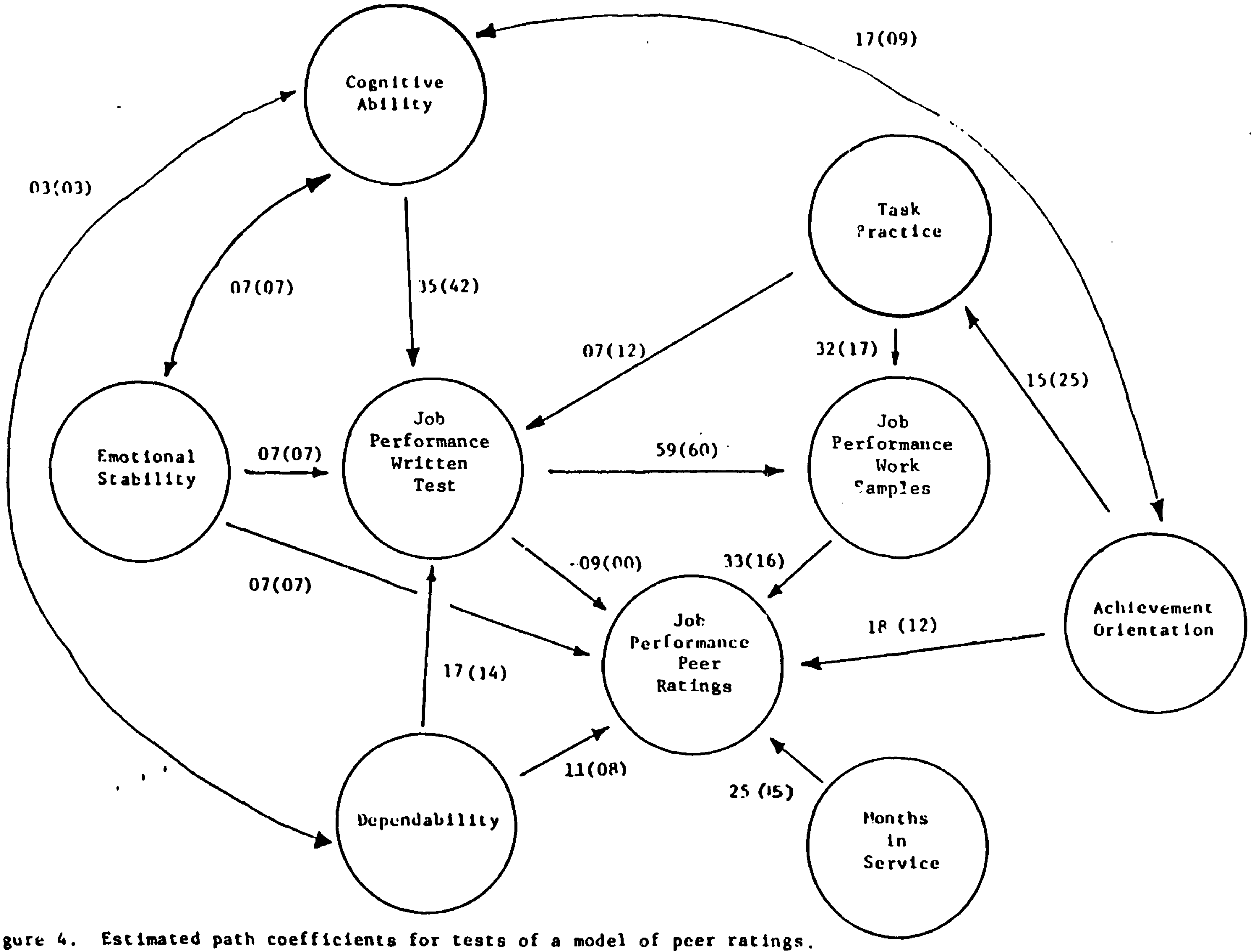


Figure 4. Estimated path coefficients for tests of a model of peer ratings.

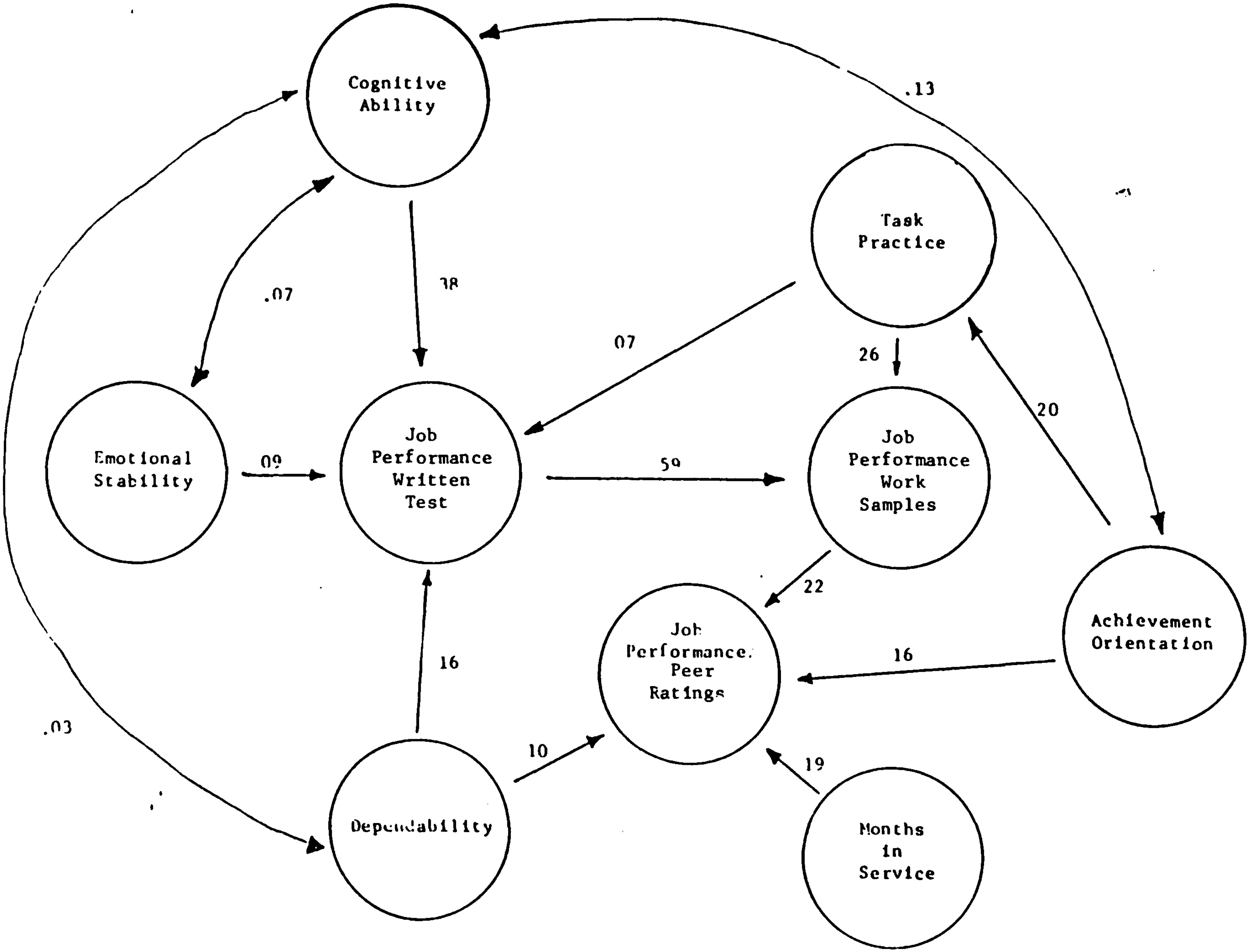


Figure 5. Model of peer rating that fit the data (n=1500).

to maintenance of knowledges and skills required to do the job. Further, soldiers with greater time in service were evaluated more favorably by their supervisors and peers. In sum, adding measures of ratee temperament and job experience to the model clearly increased the multiple correlation for the prediction of job performance ratings.

In the model, the multiple correlation for the prediction of performance ratings was higher for supervisors than peers. To some extent, supervisors seemed to be more sensitive to ratee differences in dependability and achievement orientation in rating job performance. Nevertheless, within the set of variables examined here, ratings by supervisors and peers were to a large extent influenced by similar kinds of factors. This pattern of results may be somewhat idiosyncratic to the Army environment where supervisors often work closely with their troops and might tend to have perspectives on performance similar to peers. The generality of the findings reported here should be investigated in other settings.

Work sample performance had a direct, positive impact on job performance ratings. This finding suggests that raters use information on how well the ratee can perform the tasks required by the job in making performance judgments. For the most part, job knowledge influenced ratings indirectly by contributing to work sample performance. In addition, some of the contribution of job knowledge to job performance ratings was "spurious," due to the joint influence of dependability on job knowledge and rated effectiveness.

The direct impact of job knowledge on ratings reported here was lower than expected based on previous research (Borman et al., 1985; Hunter, 1983). A close examination of past research indicates that the relative contribution of technical proficiency and job knowledge to job performance ratings may vary somewhat by job. In the 14 studies summarized by Hunter (1983), the disattenuated correlation between measures of job knowledge and overall job performance ratings varied from .08 to .63. Further, there is evidence (Hough, Gast, White & McCloy, 1986), indicating that the impact of job knowledge on overall job performance ratings may be stronger (i.e., .10 - .25) in Army combat jobs. The important point that can be made at this time is that job performance ratings by peers and supervisors are consistently related to measured performance capability.

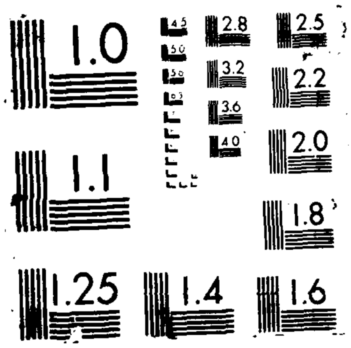
As hypothesized, job experience influenced performance on the job. Opportunities to practice job-related tasks had a direct effect on work sample performance and weaker links to measures of job knowledge. In addition, overall job experience had a direct effect on effectiveness ratings by supervisors and peers. One explanation for this finding may be that work experience contributes to rated performance by providing opportunities to aquire informal, tacit knowledge about how to stay out of trouble and manage oneself on a daily basis so as to improve productivity (Wagner & Sternberg, 1985). Other explanations are of course also possible. For example, it may be that raters are hesitant to give the very highest ratings to new, less experienced job incumbents. It should be noted that the range of job experience was relatively low in the research reported here. The planned longitudinal phase of the Project A research program should provide more definitive evidence regarding links between job experience and performance.

Ratings of job performance were shown to be related to work sample performance and job knowledge, but for the most part the different methods of performance measurement yield different results. The level of convergence across methods could be viewed as problematic and implying a lack of validity. However, the pattern of findings reported here and arguments presented elsewhere (e.g. Hanser, et al., 1985) suggest that the different methods are tapping somewhat different aspects of effectiveness on the job. Accordingly, we obtain better coverage of criterion performance by employing multiple measures, provided of course that each of the measures is tapping important facets of job performance.

Work samples and written performance tests have been referred to as maximal performance measures and capture the "can do" part of job performance. Ratings should be measuring more the motivation-related, performance-over-time aspects of performance. This is confirmed in part by the pattern of relationships between the different predictor measures and each criterion element reported here. Cognitive ability had the strongest impact on tests of job knowledge, but weaker links to ratings. Achievement orientation was more strongly related to ratings than to the work sample criterion or tests of job knowledge. Dependability had significant direct effects on both rated effectiveness and job knowledge. Importantly, the temperament scales examined here showed low correlations with cognitive ability and contribute to performance in ways not captured by tests of cognitive ability. Thus, the prediction of job performance could be increased by including such non-cognitive predictors in a composite measure with ability tests.

To sum up, the present research shed some light on factors influencing aggregated supervisor and peer ratings. The pattern of path coefficients showed that ratee differences in measured performance capability, temperament, and job experience are significant contributors to job performance ratings. Further, supervisors and peers seemed to focus on similar but not identical factots in making performance judgments. Future planned research will focus on characteristics of individual raters and rater x ratee interaction effects as po:ential contributors to job performance ratings in the model. In addition, data collected in the Project A program will allow for examination of these relationships in other Army jobs.

AD-A196 274 IMPROVING THE SELECTION CLASSIFICATION AND UTILIZATION 5/5
OF ARMY ENLISTED P.. (U) HUMAN RESOURCES RESEARCH
ORGANIZATION ALEXANDRIA VA J P CAMPBELL MAY 88
UNCLASSIFIED HUMRRO-IR-PRD-87-12-SUPPL F/G 5/9 NL


1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

REFERENCES

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bernardin, H. H., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Borman, W. C., White, L. A., Gast, I. F., & Pulakos, E. D. (August, 1985). Performance ratings as criteria: What is being measured? Paper presented at the meeting of the American Psychological Association, Los Angeles, CA.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). Development and field test of Project A task-based MOS-specific criterion measures. U.S. Army Research Institute, Technical Report, in press.

Eaton, N. K., Hanser, L. M., & Shields, J. L. (in press). Validating selection tests against job performance. In J. Zeidner (Ed.), Human Productivity Enhancement. New York: Praeger.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Guion, R. M. (1983). Comments on Hunter. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory (pp. 267-276). Jersey: Lawrence Earlbaum Associates.

Guion, R.M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. Personnel Psychology, 18, 135-164.

Hanser, L. M., Arabian, J. M., & Wise L. (October, 1985). Multi-dimensional performance measurement. Paper presented at the meeting of the Military Testing Association, San Diego, CA.

Harvey, R. J., Billings, R. S., & Nilan, K. J. (1985). Confirmatory factor analysis of the Job Diagnostic Survey: Good news and bad news. Journal of Applied Psychology, 70, 461-168.

Hough, L. M., Kamp, J. D., & Barge, B. A. (1984). Utility of temperament, biodata, and interest assessment for predicting job performance: A review and integration of the literature. Minneapolis, MN: Personnel Decisions Research Institute.

Hough, L. M. (1984). Identification and development of temperament and interest constructs and inventories for predicting job performance of Army enlisted personnel. Minneapolis: Personnel Decisions Research Institute.

Hough, L. M., & Ashworth, S. (1986). Project A Concurrent Validity data analysis: Temperament and interest predictors. Minneapolis: Personnel Decisions Research Institute.

Hough, L. M., White, L. A., Gast, I. F., & McCloy. (August, 1986). The relation of leadership and individual differences to job performance. paper presented at the meeting of the Amnerican Psychological Association, Washington, D.C.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. Performance measurement and theory (pp. 257-266). New Jersey: Lawrence Earlbaum Associates.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating results across studies. Beverly Hills: Sage Publications.

Hunter, S. E., & Hunter R. F. (1984). Validity and utility of alternative predictores of job performance. Psychological Bulletin, 96, 72-98.

Ilgen, D. R., & Feldman, J. M. (1983). Performance Appraisal: A process focus In L. Cummings & B. Staw (Eds.). Reserach in organizational behavior, Vol 3. Greenwich, CN: JAI Press.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal Analysis: Asumptions, models, and data. Beverly Hills: Sage Publications.

Joreskog, K., & Sorbom, (1979). Advances in factor analysis and structural equation models, Cambridge: Abt Books.

Joreskog, K., & Sorbom, D. (1985). LISREL: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: National Educational Resources, Inc.

Kenny, D. A. (1979). Correlation and causality. New York: Wiley.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 78-107.

Martin, C., Park, R., & Borman, C. (April, 1986). Validating a computer adaptive testing system using structural analysis. Paper presented at the American Educational Research Association, San Francisco, CA.

McDaniel, M. A. (1985). The evaluation of a causal model of job performance: Interrelationships of general mental ability, job experience, and job performance. Unpublished doctoral dissertation. George Washington University, Washington, D.C.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Schmidt, F. L., Hunter, J. E. & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample peformance, and supervisory ratings of job performance. Journal of Applied Psychology, 71, 432-439.

Smith, P. C., & Kendall, J. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Spiker, U. A., Harper, W. R., & Hayes, J. F. (1985). The effect of job experience on the maintenance proficiency of Army automatic mechanics Human Factors, 27, 301-311.

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. J. (1986). Development and field test of behaviorally anchored rating scales for nine MOS. (ARI Technical Report, in press).

Vineberg, R., & Taylor, E. N. (1972). Performance in four Army jobs by men at different aptitude (AFQT) levels: 3. The relationship of AFQT and job experience to performance. Alexandria, VA: Human Resources Research Organization.

Wagner, R.K., & Sternberg, R.J. (1985). Practical intelligence in real world pursuits: The role of tacit knowledge. Journal of Personality and Social Psychology, 49, 436-458.

A LATENT STRUCTURE MODEL OF JOB PERFORMANCE FACTORS

Lauress L. Wise
American Institutes for Research

John P. Campbell
Human Resources Research Organization

Jeffrey J. McHenry
American Institutes for Research

Lawrence M. Hanser
U.S. Army Research Institute

Presented on symposium,
"Multiple Criteria and Multiple Jobs: Will One Model Fit All?"

At the Annual Convention of the
American Psychological Association
Washington, D.C.

August 1986

The views expressed in this paper are those of the authors and do not necessarily reflect the official opinions and policies of the U.S. Army Research Institute or the Department of the Army.

A LATENT STRUCTURE MODEL OF JOB PERFORMANCE FACTORS[1]

The previous two papers have looked at some of the challenges involved in using three different methods of performance assessment to develop job performance measures for nine different jobs. The purpose of this presentation is to consider whether a single model of the latent structure of the complete criterion space will fit the data for all nine first-tour jobs.

The analysis had four major steps:

1. We first had to decide upon a basic array of criterion scores that would constitute the input to the confirmatory analysis. In their unaggregated form, there were simply too many variables to theorize about.

2. The second step was to specify a theory, or target matrix, that could be subjected to the will of LISREL.

3. The third step was to determine whether the model could be confirmed for each of the jobs, or whether the model had to be cut and spliced somewhat from job to job to fit the data adequately.

4. The final step was to examine constancies in the fit of the overall model across some or all of the MOS.

DETERMINING THE ARRAY OF CRITERION SCORES

The first step in our analyses was to identify the basic criterion scores whose structure we would analyze. As described in the companion papers, we had the following types of measures to consider:

1. Hands-on performance measures on 15 tasks that had been carefully sampled from the domain of important tasks for each job. Each hands-on test consisted of

[1] This research was funded by the Army Research Institute Contract No. MDA 903-82-C-0531. All statements expressed in this paper are those of the authors and do not necessarily express the official opinions of the U.S. Army Research Institute or the Department of the Army.

a number of "critical" steps, with each step scored pass or fail. The number of steps within a task varied widely from a half-dozen up to a maximum of 62.

2. Two paper-and-pencil tests: a job knowledge test consisting of 3 to 15 questions on each of a sample of 30 tasks (including the 15 also sampled for hands-on testing) and a school knowledge test organized around the "plan of instruction" in advanced individual (technical) training. Each test consisted of 100 to 200 items.

3. Supervisor and peer ratings of performance. The rating scales that were administered included 12 Army-wide (i.e., the scales were the same for all jobs) behavior summary scales, from 8 to 13 job-specific behavior summary scales, ratings of performance on each of the 15 tasks tested hands-on, and a 40-item combat performance prediction questionnaire. An overall rating of general effectiveness as a soldier was also obtained.

4. Performance indicators contained in official personnel records but obtained chiefly via self-report questionnaire, including such indicators as number of letters and certificates received, physical readiness test score, Articles 15 and other disciplinary actions, and M16 qualification level. File data were also used to construct a promotion rate score (relative to expected rate for a given length of service).

The administrative measures were grouped into five scales on the basis of content; no attempts were made to further reduce these scales at this point. Separate analyses were undertaken to identify smaller, more efficient variable sets within the hands-on, the written job knowledge, and the written school knowledge measures, and within each set of the ratings measures.

Reduction of the Hands-On and Written Test Variables

Initial analyses indicated that individual task scores from the hands-on and written job knowledge tests had only moderate reliability as assessed by coefficient alpha. This hindered attempts to uncover the true relationships among the different task scores. Consequently, tasks were grouped into "functional categories" on the basis of similarity of task content. The 30 tasks sampled for each job were clustered into 8 to 15 functional categories. Each of the school knowledge items was similarly mapped into a specific functional category.

Ten of the functional categories were common to some or all of the jobs (e.g., first aid, basic weapons, field techniques). Each job, except Infantryman, also had two to five functional performance categories that were unique. Figure 1 shows the different functional categories used for each of the nine jobs.

Next, scores were computed for each functional category within each of the three sets of measures. For the hands-on test, the functional category score was the mean percent of steps passed across all of the tasks assigned to that category. For the job knowledge test and the school knowledge test, the functional category score was the percent of items within that category that were answered correctly.

After category scores were computed, they were factor analyzed. Separate factor analyses were executed for each type of measure within each job. There were several common features in the results of these analyses for the different jobs. First, the unique functional categories for each job tended to load on different factors than the common functional categories. Second, the factors that emerged from the common functional categories tended to be fairly similar across the nine different jobs and across the three methods. Some of the categories were not sampled in one or more of the tests for some jobs, so differences in the common category factors were inevitable.

Using the empirical factor analysis to guide us, we adopted the following set of content categories:

1. Basic Soldiering Skills (field techniques, weapons, navigate, customs and laws)

2. Safety/Survival (first aid, nuclear-biological-chemical safety)

3. Communications (radio operation)

4. Vehicle Maintenance

5. Identify Friendly/Enemy Aircraft and Vehicles

6. Technical Skills (specific to the job)

Figure 1

Functional Categories by Job and Method

	11B			13B			19E			31C			63B			64C			71L			91A			95B		
Cluster Number and Name	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK	HO	JK	SK
1. First Aid	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2. Navigate		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3. NBC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x
4. Weapons	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x		x	x	x
5. Field Techniques	x	x	x	x	x	x		x	x		x	x		x	x		x	x		x	x		x	x	x	x	x
6. Communication	x	x	x	x	x	x	x	x	x	x	x				x			x			x				x	x	x
7. ID Target		x	x		x			x	x		x	x					x						x			x	x
8. Customs and Laws			x		x	x		x	x		x	x	x	x	x		x	x		x	x			x			x
9. Antitank/Antiair Weapons		x	x			x						x						x			x						x
11. Drive (Operate and Maintain)	x		x			x				x	x	x			x	x	x	x			x		x	x	x	x	x
14. Prepare/Operate/Maintain Howitzer and Ammunition				x	x	x																					
15. Operate Howitzer Sight/Alignment Device				x	x	x																					
16. Preventive Maintenance															x			x									x
17. Tank Operations							x	x	x																		
18. Tank Gunnery							x	x	x																		
20. Generators										x	x																
21. TTY Station and Net Operators										x	x																
22. Maintain TTY Electronic Equipment											x	x															
23. Operate TTY Electronic Equipment										x	x																
24. Install TTY Electronic Equipment										x	x																
25. Electrical System													x	x	x												
26. Brake/Steering/Suspension System													x	x	x												
27. Vehicle Operation and Recovery													x	x	x			x									x
28. Fuel/Cooling/Lubricating													x	x	x												
29. Forms/Files Management																			x	x	x						
30. Supervision/Coordination																				x	x						
31. Correspondence																			x	x	x						
32. Classified Material																			x	x	x						
33. Clinic/Ward Treatment and Care																						x	x	x			
34. Clinic/Ward Housekeeping																						x	x	x			
35. Clinic/Ward Management																							x	x			
36. General Medical Knowledge																								x			
37. Responding To Alarms																										x	x
38. Conduct MP Procedures																										x	x
39. Patrol Duties																									x	x	x
93. Power Train and Clutch													x	x													

The resultant categories reflected theoretical concerns more than a strict adherence to empirical findings. The tasks sorted into each category were judged similar in content and perhaps also in the knowledge and skills required.

Reduction of the Rating Variables

The individual rating scales were, for the most part, highly reliable. No combining was required to achieve adequate stability. The different scales were, however, highly intercorrelated. Further reduction in the number of scales was aimed at reducing the redundancy and colinearity among these measures.

Empirical factor analyses of the Army-wide rating scales suggested three factors. These were:

1. Effort/Leadership, including effort and competence in performing job tasks, leadership, and self-development.

2. Maintaining Personal Discipline, including self-control, integrity, and following regulations.

3. Fitness and Appearance, including physical fitness and maintaining proper military bearing and appearance.

Similar factor analyses were reviewed for the job-specific behavioral summary scales for each job. Two factors were identified based on these results. The first consisted of those aspects of job performance that were central to the specific technical content of each job. The second factor included the remaining, less central job performance components. Again the final formulation of factors was based on a combination of empirical and judgmental considerations.

Some analyses of the task ratings were also conducted. In general, these scales were less reliable than either the Army-wide or the job-specific behavioral summary scales. Supervisors and peers often reported that they had never had an opportunity to observe their ratees' performance on many of the tasks, leading to a significant missing data problem. Because of these problems, and because we believed that supervisors and peers were able to evaluate ratees' performance adequately using the other sets of rating scales that had been developed, the task ratings were dropped from the present analyses.

The individual items in the combat performance prediction battery also were subjected to an empirical factor analysis. Two factors emerged. The first factor consisted of items depicting exemplary effort, skill, or courage under stressful conditions. The second factor consisted of negatively worded items portraying failure to follow instructions and lack of discipline under stressful conditions.

The Final Array

The final array of variables for each job consisted of:

- 2-5 hands-on content category scores
- 2-6 job knowledge content category scores
- 2-6 school knowledge content category scores
- 3 Army-wide rating factors
- 2 job-specific rating factors
- 2 combat performance prediction rating factors
- 1 overall effectiveness rating
- 5 administrative measures scale scores

Tables 1 through 9 show the means, standard deviations, and intercorrelations among these variables for each of the nine jobs.

BUILDING THE TARGET MODEL

The next step was to build a target model of job performance that could be tested for goodness of fit within each of our nine jobs. Campbell and Harris (1985) developed an initial model of performance constructs for entry-level enlisted specialties, shown in Figure 2. The correlation matrices shown in Tables 1 through 9 were each subjected to an empirical factor analysis to suggest possible modifications to the original model.

Several consistent results were observed in the different factor analyses. First was the general dominance of "method" factors, specifically one factor for the ratings and one for the written tests. The evidence for a "hands-on" method factor was less compelling, perhaps because there were

TABLE 1

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 11B: INFANTRY

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	Overall Rating	4.60	0.90	.	90	74	68	77	85	75	65	23	12	17	-35	36	26	14	4	35	25	11	10	33	19	18	12	14
2	Eff/Ldr Rating	4.41	0.82	90	.	74	65	80	88	80	67	24	8	13	-30	36	30	12	5	36	27	10	13	33	20	20	9	17
3	Discipline Rtng	4.50	0.87	74	74	.	49	55	71	63	66	13	3	7	-39	31	16	10	3	30	22	6	8	24	13	13	5	13
4	Fitness Rating	4.86	0.89	68	65	49	.	59	66	52	45	17	27	9	-24	22	10	9	-1	10	10	-2	-4	13	6	6	1	1
5	Job-Spec Tech	32.98	4.58	77	80	55	59	.	86	75	58	23	15	17	-20	22	27	15	5	35	22	12	10	36	21	23	9	16
6	Job-Spec Other	22.67	3.66	85	88	71	66	86	.	80	67	25	8	14	-28	32	23	10	6	35	26	12	12	33	17	22	11	17
7	Combat Exmplry	9.02	1.49	75	80	63	52	75	80	.	75	24	8	13	-31	29	28	12	7	37	25	9	16	34	22	23	9	19
8	Combat Problems	10.03	1.64	65	67	66	45	58	67	75	.	14	8	6	-33	27	20	7	-1	36	24	9	15	31	21	18	8	14
9	Awards & Certs	3.33	2.18	23	24	13	17	23	25	24	14	.	15	20	-2	4	13	6	-1	14	15	-0	13	9	9	5	4	12
10	Phys. Readiness	273.44	28.00	12	8	3	27	15	8	8	8	15	.	11	2	-6	1	-7	-9	0	5	-7	-0	8	-2	-1	-4	-6
11	M16 Qualific.	2.74	0.57	17	13	7	9	17	14	13	6	20	11	.	1	1	13	6	-0	10	2	3	0	14	10	5	3	6
12	Articles 15	0.39	0.85	-35	-30	-39	-24	-20	-28	-31	-33	-2	2	1	.	-45	-10	-1	-6	-10	-9	-8	-6	-10	-1	-9	0	-5
13	Promotion Rate	0.03	0.68	36	36	31	22	22	32	29	27	4	-6	1	-45	.	16	7	7	19	17	12	10	18	14	12	11	17
14	HO Basic	50.50	10.06	26	30	16	10	27	23	28	20	13	1	13	-10	16	.	15	6	44	30	13	27	40	24	20	16	30
15	HO Safety	22.67	3.41	14	12	10	9	15	10	12	7	6	-7	6	-1	7	15	.	2	16	8	1	8	16	7	3	3	4
16	HO Comm	13.15	1.53	4	5	3	-1	5	6	7	-1	-1	-9	-0	-6	7	6	2	.	4	6	-1	-3	0	4	6	2	-1
17	JK Basic	50.93	9.71	35	36	30	10	35	35	37	36	14	0	10	-10	19	44	16	4	.	68	40	42	65	50	40	30	35
18	JK Safety	20.02	4.31	25	27	22	10	22	26	25	24	15	5	2	-9	17	30	8	6	68	.	23	26	47	41	32	25	20
19	JK Comm	4.37	1.47	11	10	6	-2	12	12	9	9	-0	-7	3	-6	12	13	1	-1	40	23	.	16	26	25	19	14	16
20	JK Identify	8.25	2.24	10	13	8	-4	10	12	16	15	13	-0	0	-6	10	27	8	-3	42	26	16	.	31	24	18	16	37
21	SK Basic	72.87	14.89	33	33	24	13	36	33	34	31	9	8	14	-10	18	40	16	0	65	47	26	31	.	63	60	44	43
22	SK Safety	9.51	2.12	19	20	13	6	21	17	22	21	9	-2	10	-1	14	24	7	4	50	41	25	24	63	.	45	34	26
23	SK Comm	5.68	1.67	18	20	13	6	23	22	23	18	5	-1	5	-9	12	20	3	6	40	32	19	18	60	45	.	40	31
24	SK Vehicle	0.78	0.42	12	9	5	1	9	11	9	8	4	-4	3	0	11	16	3	2	30	25	14	16	44	34	40	.	21
25	SK Identify	2.30	1.16	14	17	13	1	16	17	19	14	12	-6	6	-5	17	30	4	-1	35	20	18	37	43	26	31	21	.

N= 503

TABLE 2

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 13B: CANNON CREWMAN

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	Overall Rating	4.59	0.79	.	86	71	61	62	72	73	61	11	10	5	-25	30	20	19	17	6	26	18	14	8	3	24	15	12	8	9
2	Eff/Ldr Rating	4.43	0.76	86	.	75	62	65	74	78	61	14	6	1	-23	25	27	25	14	9	32	20	15	11	5	30	20	15	5	13
3	Discipline Rtng	4.61	0.78	71	75	.	51	53	60	63	60	-0	-4	-1	-20	26	12	9	12	4	22	16	15	4	3	18	14	14	6	16
4	Fitness Rating	4.95	0.82	61	62	51	.	47	53	51	39	7	23	-1	-25	16	8	4	0	3	5	-1	-1	1	-2	4	-4	-1	-4	-8
5	Job-Spec Tech	23.59	3.55	62	65	53	47	.	80	60	39	11	10	1	-2	10	35	18	9	-1	25	10	10	17	8	24	8	12	6	4
6	Job-Spec Other	23.90	3.08	72	74	60	53	80	.	66	49	6	5	-4	-9	18	25	18	8	1	29	18	15	13	6	26	14	16	4	8
7	Combat Exmplry	9.00	1.44	73	78	63	51	60	66	.	63	14	10	3	-15	23	20	23	13	3	22	16	13	6	8	23	12	7	-1	1
8	Combat Problems	9.92	1.56	61	61	60	39	39	49	63	.	8	7	-3	-16	26	14	16	8	12	19	17	10	14	8	15	14	9	5	8
9	Awards & Certs	2.58	1.82	11	14	-0	7	11	6	14	8	.	12	18	0	8	15	19	15	-1	11	10	6	5	8	11	6	5	8	2
10	Phys. Readiness	261.74	32.70	10	6	-4	23	10	5	10	7	12	.	11	-3	-2	7	2	-7	8	-8	-8	-10	5	4	-0	-8	-10	-12	-15
11	M16 Qualific.	2.25	0.69	5	1	-1	-1	1	-4	3	-3	18	11	.	6	1	7	8	12	-3	-4	-5	-6	7	-3	-3	-7	0	3	-3
12	Articles 15	0.46	1.03	-25	-23	-20	-25	-2	-9	-15	-16	0	-3	6	.	-31	-0	-4	-5	-5	-7	-10	-12	-7	1	-5	-6	-2	-5	-3
13	Promotion Rate	0.01	0.63	30	25	26	16	10	18	23	26	8	-2	1	-31	.	6	10	10	3	10	6	5	5	-1	2	5	-2	10	7
14	HO Tech.	50.71	9.94	20	27	12	8	35	25	20	14	15	7	7	-0	6	.	47	20	11	33	13	7	10	12	36	18	20	11	9
15	HO Basic	48.50	13.00	19	25	9	4	18	18	23	16	19	2	8	-4	10	47	.	21	8	42	38	20	9	15	40	25	17	15	9
16	HO Safety	40.16	6.23	17	14	12	0	9	8	13	8	15	-7	12	-5	10	20	21	.	11	24	14	11	9	3	25	20	16	11	24
17	HO Comm	10.60	1.59	6	9	4	3	-1	1	3	12	-1	8	-3	-5	3	11	8	11	.	1	1	-2	6	5	7	5	-1	1	3
18	JK Tech.	50.67	9.94	26	32	22	5	25	29	22	19	11	-8	-4	-7	10	33	42	24	1	.	58	54	21	20	64	52	41	37	35
19	JK Basic	31.91	5.78	18	20	16	-1	10	18	16	17	10	-8	-5	-10	6	13	38	14	1	58	.	55	14	23	52	49	38	35	27
20	JK Safety	23.58	4.43	14	15	15	-1	10	15	13	10	6	-10	-6	-12	5	7	20	11	-2	54	55	.	10	21	41	38	35	26	27
21	JK Comm	1.12	0.68	8	11	4	1	17	13	6	14	5	5	7	-7	5	10	9	9	6	21	14	10	.	13	19	13	16	14	11
22	JK Identify	7.12	2.25	3	5	3	-2	8	6	8	8	8	4	-3	1	-1	12	15	3	5	20	23	21	13	.	20	21	25	10	8
23	SK Tech.	50.82	9.84	24	30	18	4	24	26	23	15	11	-0	-3	-5	2	36	40	25	7	64	52	41	19	20	.	63	47	38	40
24	SK Basic	23.17	5.27	15	20	14	-4	8	14	12	14	6	-8	-7	-6	5	18	25	20	5	52	49	36	13	21	63	.	51	40	52
25	SK Safety	8.44	2.12	12	15	14	-1	12	16	7	9	5	-10	0	-2	-2	20	17	18	-1	41	38	35	16	25	47	51	.	28	36
26	SK Comm	3.55	1.21	8	5	6	-4	6	4	-1	5	8	-12	3	-5	10	11	15	11	1	37	35	26	14	10	33	40	28	.	32
27	SK Vehicle	2.75	1.07	9	13	16	-8	4	8	1	8	2	-15	-3	-3	7	9	9	24	3	35	27	27	11	8	40	52	36	32	.

N= 401

TABLE 3

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 19E: ARMOR CREWMAN

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Overall Rating	4.62	0.78	.	84	72	58	72	53	69	61	12	20	8	-37	41	12	15	16	4	25	23	22	19	10	27	26	18	22	18	7
2	Eff/Ldr Rating	4.38	0.74	84	.	68	55	76	50	80	65	16	21	8	-32	41	17	26	19	17	31	34	31	27	14	33	32	23	22	23	11
3	Discipline Rtng	4.50	0.83	72	68	.	45	53	41	55	64	-1	12	-14	-35	38	6	15	14	2	21	23	18	17	6	27	18	8	22	14	-2
4	Fitness Rating	4.76	0.82	58	55	45	.	44	39	43	36	10	43	-0	-19	28	8	2	16	-3	1	5	10	5	-4	0	-0	-0	4	2	2
5	Job-Spec Tech	23.19	3.20	72	76	53	44	.	75	71	55	10	14	17	-31	34	23	17	19	13	25	23	26	19	8	27	25	18	15	20	6
6	Job-Spec Other	14.71	1.89	53	50	41	39	75	.	50	41	9	13	13	-16	19	15	9	13	2	8	7	12	2	4	15	12	8	9	9	1
7	Combat Exmplry	8.88	1.36	69	80	55	43	71	50	.	63	15	18	8	-32	34	15	27	15	14	20	23	19	20	7	22	25	19	10	18	2
8	Combat Problems	9.80	1.47	61	65	64	36	55	41	63	.	-1	7	4	-31	29	13	22	13	6	24	18	21	13	8	24	18	17	15	16	-1
9	Awards & Certs	2.52	1.60	12	16	-1	10	10	9	15	-1	.	15	19	-7	13	6	4	-3	13	5	7	-0	10	-2	12	12	3	4	8	8
10	Phys. Readiness	249.41	27.11	20	21	12	43	14	13	18	7	15	.	-1	-10	10	-3	-3	4	2	-6	0	-6	4	1	-4	1	2	-2	2	-7
11	M16 Qualific.	2.40	0.68	8	8	-14	-0	17	13	8	4	19	-1	.	14	-1	7	7	3	10	11	12	13	17	31	10	6	12	2	16	-1
12	Articles 15	0.35	0.77	-37	-32	-35	-19	-31	-18	-32	-31	-7	-10	14	.	-43	-9	-8	-16	1	-13	-17	-17	-7	1	-19	-13	-13	-0	-7	-6
13	Promotion Rate	0.03	0.58	41	41	38	28	34	19	34	29	13	10	-1	-43	.	10	7	15	12	14	24	28	21	2	17	22	18	6	15	1
14	HO Tech.	50.00	9.99	12	17	6	8	23	15	15	13	6	-3	7	-9	10	.	18	24	20	36	27	27	13	18	23	18	9	2	19	0
15	HO Basic	38.16	2.48	15	26	15	2	17	9	27	22	4	-3	7	-8	7	18	.	21	23	30	32	25	21	18	21	25	11	4	19	-0
16	HO Safety	21.85	2.95	16	19	14	16	19	13	15	13	-3	4	3	-16	15	24	21	.	14	22	18	18	10	6	15	13	5	5	17	6
17	HO Comm	28.55	7.59	4	17	2	-3	13	2	14	6	13	2	10	1	12	20	23	14	.	23	28	25	32	11	20	23	13	3	23	3
18	JK Tech.	50.00	9.99	25	31	21	1	25	8	20	24	5	-6	11	-13	14	36	30	22	23	.	60	52	45	34	64	60	44	38	42	7
19	JK Basic	42.16	7.28	23	34	23	5	23	7	23	18	7	0	12	-17	24	27	32	18	28	60	.	65	53	30	65	67	46	41	43	6
20	JK Safety	21.19	4.10	22	31	18	10	26	12	19	21	-0	-6	13	-17	28	27	25	16	25	52	65	.	44	34	46	51	37	26	33	5
21	JK Comm	11.33	3.59	19	27	17	5	19	2	20	13	10	4	17	-7	21	13	21	10	32	45	53	44	.	16	45	51	34	30	24	2
22	JK Identify	10.05	1.78	10	14	6	-4	8	4	7	8	-2	1	31	1	2	18	18	6	11	34	30	34	16	.	24	28	22	18	37	3
23	SK Tech.	54.54	9.66	27	33	27	0	27	15	22	24	12	-4	10	-19	17	23	21	15	20	64	65	46	45	24	.	75	53	59	48	21
24	SK Basic	34.94	8.44	26	32	18	-0	25	12	25	18	12	1	6	-13	22	18	25	13	23	60	67	51	51	23	75	.	68	47	47	12
25	SK Safety	8.18	2.14	18	23	8	-0	18	8	19	17	3	2	12	-13	18	9	11	5	13	44	46	37	34	22	53	68	.	38	33	4
26	SK Comm	7.59	1.80	22	22	22	4	15	9	10	15	4	-2	2	-0	6	2	4	5	3	38	41	26	30	18	59	47	38	.	24	14
27	SK Vehicle	0.54	0.50	7	11	-2	2	6	1	2	-1	8	-7	-1	-6	1	0	-0	6	3	7	6	5	2	3	21	12	4	14	9	.
28	SK Identify	3.01	0.96	18	23	14	2	20	9	18	16	8	2	16	-7	15	19	19	17	23	42	43	33	24	37	48	47	33	24	.	9

N= 335

TABLE 4

JOB PERFORMANCE MEASURE SUMMARY STATISTICS FOR 31C: SINGLE CHANNEL RADIO OPERATOR

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	Overall Rating	4.73	0.79	.	83	73	64	74	66	66	66	17	11	2	-31	30	20	24	15	15	-2	24	17	9	14	3	13	19	10	14	2	-2
2	Eff/Ldr Rating	4.48	0.72	83	.	68	57	81	71	68	63	18	12	7	-31	30	24	21	21	15	2	30	28	14	16	6	13	23	12	20	12	4
3	Discipline Rtng	4.64	0.88	73	68	.	52	54	58	53	60	4	4	-11	-32	26	10	14	7	10	-1	20	15	4	15	6	7	9	-4	10	-2	-8
4	Fitness Rating	5.05	0.88	64	57	52	.	47	40	42	42	11	34	-6	-25	24	12	8	10	4	-2	1	2	0	8	-4	6	-5	-6	-2	-9	-12
5	Job-Spec Tech	14.27	2.01	74	81	54	47	.	76	66	57	14	4	5	-16	22	20	24	20	11	-1	29	30	16	15	8	15	19	15	16	13	-1
6	Job-Spec Other	14.37	2.09	66	71	58	40	76	.	54	48	3	-3	-3	-17	22	11	18	15	1	0	17	22	8	9	2	5	5	2	9	3	-9
7	Combat Exmplry	9.09	1.54	66	68	53	42	66	54	.	77	11	1	5	-21	17	6	13	18	23	-7	26	30	15	19	11	12	18	9	14	10	7
8	Combat Problems	10.47	1.71	66	63	60	42	57	48	77	.	9	-1	-2	-22	14	4	16	11	15	-0	22	24	3	14	-1	5	15	5	9	0	-3
9	Awards & Certs	2.16	1.75	17	18	4	11	14	3	11	9	.	23	10	2	12	9	12	6	3	2	10	10	11	-0	-5	8	8	12	4	4	6
10	Phys. Readiness	259.54	29.59	11	12	4	34	4	-3	1	-1	23	.	4	-11	4	1	-10	0	1	-6	-4	-8	4	1	3	-8	-4	-0	1	-13	-5
11	M16 Qualific.	2.16	0.77	2	7	-11	-6	5	-3	5	-2	10	4	.	4	3	4	5	10	7	5	7	10	8	-4	-6	5	9	4	4	11	10
12	Articles 15	0.34	0.84	-31	-31	-32	-25	-16	-17	-21	-22	2	-11	4	.	-34	-9	-3	-7	-12	-3	-16	-9	-13	-20	-10	-3	-11	-4	-12	-4	-3
13	Promotion Rate	-0.02	0.56	30	30	26	24	22	22	17	14	12	4	3	-34	.	8	12	21	9	5	18	17	10	19	13	12	13	15	12	4	-0
14	HO Tech.	78.44	9.49	20	24	10	12	20	11	6	4	9	1	4	-9	8	.	25	25	28	9	42	21	23	21	22	15	39	21	24	9	8
15	HO Basic	21.25	3.84	24	21	14	8	24	18	13	16	12	-10	5	-3	12	25	.	18	27	8	31	31	18	15	5	21	27	24	27	10	15
16	HO Safety	20.15	3.99	15	21	7	10	20	15	18	11	6	0	10	-7	21	25	18	.	23	16	10	21	13	9	6	8	11	10	19	4	9
17	HO Comm	16.73	6.59	15	15	10	4	11	1	23	15	3	1	7	-12	9	28	27	23	.	1	34	29	21	38	21	23	26	17	11	5	25
18	HO Vehicle	11.73	1.31	-2	2	-1	-2	-1	0	-7	-0	2	-6	5	-3	5	9	8	16	1	.	11	9	10	2	-6	7	22	16	14	12	11
19	JK Tech.	57.16	11.68	24	30	20	1	29	17	26	22	10	-4	7	-16	18	42	31	10	34	11	.	60	59	60	37	33	72	49	50	44	25
20	JK Basic	22.12	4.61	17	29	15	2	30	22	30	24	10	-8	10	-9	17	21	31	21	29	9	60	.	58	50	23	31	49	42	43	40	20
21	JK Safety	23.31	4.63	9	14	4	0	16	8	15	3	11	4	8	-13	10	23	18	13	21	10	59	58	.	50	28	30	44	40	48	36	24
22	JK Comm	10.12	2.74	14	16	15	8	15	9	19	14	-0	1	-4	-20	19	21	15	9	38	2	60	50	50	.	32	19	44	36	36	27	19
23	JK Vehicle	4.54	1.82	3	6	6	-4	8	2	11	-1	-5	3	-6	-10	13	22	5	6	21	-6	37	23	28	32	.	17	20	14	16	13	11
24	JK Identify	6.72	2.13	13	13	7	6	15	5	12	5	8	-6	5	-3	12	15	21	8	23	7	33	31	30	19	17	.	27	28	21	11	40
25	SK Tech.	77.87	15.43	19	23	9	-5	19	5	18	15	8	-4	9	-11	13	39	27	11	26	22	72	49	44	44	20	27	.	62	58	48	29
26	SK Basic	10.95	2.74	10	12	-4	-8	15	2	9	5	12	-0	4	-4	15	21	24	10	17	16	49	42	40	36	14	28	62	.	56	42	26
27	SK Safety	11.08	2.81	14	20	10	-2	16	9	14	9	4	1	4	-12	12	24	27	19	11	14	50	43	48	36	16	21	58	56	.	41	27
28	SK Vehicle	3.83	1.84	2	12	-2	-9	13	3	10	0	4	-13	11	-4	4	9	10	4	5	12	44	40	36	27	13	11	48	42	41	.	22
29	SK Identify	1.16	0.73	-2	4	-8	-12	-1	-9	7	-3	6	-5	10	-3	-0	8	15	9	25	11	25	20	24	19	11	40	29	26	27	22	

N= 239

TABLE 5

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 63B: LIGHT WEIGHT VEHICLE MECHANIC

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	Overall Rating	4.55	0.84	.	86	75	57	75	75	68	65	20	7	-4	-24	24	11	-1	5	10	20	15	22	11	21	22	21	15	19
2	Eff/Ldr Rating	4.31	0.83	86	.	75	50	84	78	69	66	21	1	-5	-23	23	19	-1	3	12	23	16	27	18	26	22	19	14	22
3	Discipline Rtng	4.54	0.88	75	75	.	51	63	65	59	66	15	2	-8	-27	26	10	-5	7	9	11	5	19	3	14	23	20	13	14
4	Fitness Rating	4.82	0.86	57	50	51	.	38	49	44	41	13	31	2	-20	20	-2	-2	8	7	-0	2	8	-0	-2	16	14	13	8
5	Job-Spec Tech	22.42	4.10	75	84	63	38	.	78	65	57	21	-1	-5	-16	16	23	1	3	13	28	21	26	19	37	21	18	6	28
6	Job-Spec Other	23.19	3.52	75	78	65	49	78	.	68	55	18	5	-8	-18	17	12	4	4	12	18	17	22	13	21	18	18	9	20
7	Combat Exmplry	8.67	1.61	68	69	59	44	65	68	.	69	14	4	-7	-16	17	13	0	9	9	16	11	23	8	20	18	14	8	13
8	Combat Problems	9.92	1.86	65	66	66	41	57	55	69	.	14	-0	-6	-20	27	10	-3	4	9	17	11	20	7	19	21	18	18	18
9	Awards & Certs	2.31	1.81	20	21	15	13	21	18	14	14	.	4	2	-11	7	11	-5	-0	7	7	2	12	11	13	14	10	8	18
10	Phys. Readiness	255.47	31.93	7	1	2	31	-1	5	4	-0	4	.	10	-10	15	1	8	3	-1	-7	-12	-2	-9	-10	1	0	-3	-4
11	M16 Qualific.	2.19	0.73	-4	-5	-8	2	-5	-8	-7	-6	2	10	.	1	-9	-2	5	-4	-0	-6	3	3	2	-2	-2	2	-0	4
12	Articles 15	0.37	0.85	-24	-23	-27	-20	-16	-18	-16	-20	-11	-10	1	.	-36	-3	-2	-2	-4	-7	-5	-6	-0	-6	-11	-7	-13	-8
13	Promotion Rate	0.04	0.52	24	23	26	20	16	17	17	27	7	15	-9	-36	.	-5	-4	-2	-1	13	9	4	8	13	16	15	8	13
14	HO Tech.	110.11	6.84	11	19	10	-2	23	12	13	10	11	1	-2	-3	-5	.	8	6	18	33	23	19	22	37	19	16	4	24
15	HO Basic	34.96	4.09	-1	-1	-5	-2	1	4	0	-3	-5	8	5	-2	-4	8	.	10	7	6	12	14	12	10	7	15	-1	14
16	HO Safety	21.92	3.25	5	3	7	8	3	4	9	4	-0	3	-4	-2	-2	6	10	.	2	2	5	18	1	1	2	7	-7	-0
17	HO Vehicle	11.22	1.84	10	12	9	7	13	12	9	9	7	-1	-0	-4	-1	18	7	2	.	15	6	4	11	17	6	6	2	13
18	JK Tech.	66.61	11.93	20	23	11	-0	28	18	16	17	7	-7	-6	-7	13	33	6	2	15	.	62	47	62	67	50	39	36	59
19	JK Basic	24.36	4.69	15	16	5	2	21	17	11	11	2	-12	3	-5	9	23	12	5	6	62	.	45	44	47	41	36	22	44
20	JK Safety	18.91	3.05	22	27	19	8	26	22	23	20	12	-2	3	-6	4	19	14	18	4	47	45	.	38	40	36	33	20	39
21	JK Vehicle	15.81	4.03	11	18	3	-0	19	13	8	7	11	-9	2	-0	8	22	12	1	11	62	44	38	.	56	37	31	24	49
22	SK Tech.	56.00	12.69	21	26	14	-2	37	21	20	19	13	-10	-2	-6	13	37	10	1	17	67	47	40	56	.	52	47	30	69
23	SK Basic	16.56	4.24	22	22	23	16	21	18	18	21	14	1	-2	-11	16	19	7	2	6	50	41	36	37	52	.	61	50	56
24	SK Safety	6.02	1.74	21	19	20	14	18	18	14	18	10	0	2	-7	15	16	15	7	6	39	36	33	31	47	61	.	39	50
25	SK Comm	0.90	0.30	15	14	13	13	6	9	6	18	8	-3	-0	-13	8	4	-1	-7	2	36	22	20	24	30	50	39	.	39
26	SK Vehicle	24.10	5.54	19	22	14	8	28	20	13	18	18	-4	4	-8	13	24	14	-0	13	59	44	39	49	69	56	50	39	.

N= 403

TABLE 6

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 64C: MOTOR TRANSPORT OPERATOR

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Overall Rating	4.52	0.76	.	86	78	63	72	59	68	58	13	11	4	-30	33	8	20	15	16	17	19	3	13	12	16	14
2	Eff/Ldr Rating	4.36	0.75	86	.	77	59	78	69	74	58	17	9	6	-25	31	16	20	18	23	22	26	7	21	17	15	21
3	Discipline Rtng	4.53	0.81	78	77	.	52	67	51	58	54	10	3	-2	-29	35	4	14	14	15	15	19	1	16	12	15	16
4	Fitness Rating	4.74	0.87	63	59	52	.	54	39	46	35	3	28	3	-20	21	-2	8	14	5	6	4	2	5	6	7	-2
5	Job-Spec Tech	29.61	3.76	72	78	67	54	.	78	65	52	13	6	7	-21	25	9	16	19	17	20	19	5	16	17	12	15
6	Job-Spec Other	17.79	2.52	59	69	51	39	78	.	63	41	18	4	13	-15	19	12	11	16	17	16	19	4	13	17	7	14
7	Combat Exmplry	8.80	1.45	68	74	58	46	65	63	.	65	12	6	11	-21	22	20	19	16	20	15	20	5	18	8	10	15
8	Combat Problems	9.50	1.63	58	58	54	35	52	41	65	.	8	-3	2	-24	26	12	15	10	16	17	22	7	15	14	20	19
9	Awards & Certs	3.12	2.08	13	17	10	3	13	18	12	8	.	6	11	5	12	8	4	5	-3	-2	1	6	3	4	-1	2
10	Phys. Readiness	248.48	37.70	11	9	3	28	6	4	6	-3	6	.	3	-6	-1	-1	3	2	-4	-4	-4	2	-2	-5	0	-8
11	M16 Qualific.	2.09	0.75	4	6	-2	3	7	13	11	2	11	3	.	4	-5	9	13	5	7	5	3	-0	-1	-3	1	-1
12	Articles 15	0.46	0.98	-30	-25	-29	-20	-21	-15	-21	-24	5	-6	4	.	-36	-1	-11	-11	-7	-13	-12	0	-5	-8	-12	-4
13	Promotion Rate	-0.01	0.57	33	31	35	21	25	19	22	26	12	-1	-5	-36	.	10	9	10	9	12	11	5	11	9	8	11
14	HO Basic	43.44	10.16	8	16	4	-2	9	12	20	12	8	-1	9	-1	10	.	29	10	44	31	30	7	28	21	6	32
15	HO Safety	83.73	9.84	20	20	14	8	16	11	19	15	4	3	13	-11	9	29	.	14	27	31	24	4	24	19	14	24
16	HO Vehicle	33.30	4.19	15	18	14	14	19	16	16	10	5	2	5	-11	10	10	14	.	5	8	15	3	10	11	1	11
17	JK Basic	27.38	5.82	16	23	15	5	17	17	20	16	-3	-4	7	-7	9	44	27	5	.	67	54	10	47	29	20	49
18	JK Safety	33.42	5.42	17	22	15	6	20	16	15	17	-2	-4	5	-13	12	31	31	8	67	.	49	4	42	47	23	49
19	JK Vehicle	35.40	7.70	19	26	19	4	19	19	20	22	1	-4	3	-12	11	30	24	15	54	49	.	11	49	40	27	55
20	JK Identify	2.15	1.41	3	7	1	2	5	4	5	7	6	2	-0	0	5	7	4	3	10	4	11	.	17	10	-2	13
21	SK Basic	16.41	4.36	13	21	16	5	16	13	18	15	3	-2	-1	-5	11	28	24	10	47	42	49	17	.	56	43	68
22	SK Safety	6.44	1.93	12	17	12	6	17	17	8	14	4	-5	-3	-8	9	21	19	11	39	47	40	10	56	.	36	59
23	SK Comm	0.89	0.32	16	15	15	7	12	7	10	20	-1	0	1	-12	8	6	14	1	20	23	27	-2	43	36	.	37
24	SK Vehicle	55.72	10.07	14	21	16	-2	15	14	15	19	2	-8	-1	-4	11	32	24	11	49	49	55	13	68	59	37	.

N= 477

TABLE 7

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 71L: ADMINISTRATIVE SPECIALIST

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Overall Rating	4.92	0.85	.	83	71	57	72	63	63	59	20	24	4	-23	20	17	14	3	22	15	17	21	13	11	5	10
2	Eff/Ldr Rating	4.64	0.78	83	.	73	56	73	65	70	60	21	19	2	-19	19	25	14	2	29	17	18	28	17	9	7	11
3	Discipline Rtng	5.01	0.88	71	73	.	47	63	55	58	58	13	13	4	-27	19	20	10	-3	22	15	11	20	7	8	1	4
4	Fitness Rating	5.23	0.89	57	56	47	.	40	39	55	49	20	35	5	-23	20	3	7	-3	1	2	2	-1	0	-5	0	-2
5	Job-Spec Tech	19.88	2.73	72	73	63	40	.	76	54	50	8	7	-5	-21	21	24	8	-2	28	16	16	28	10	9	6	7
6	Job-Spec Other	18.57	3.13	63	65	55	39	76	.	50	46	10	13	-1	-21	17	22	13	1	22	15	16	26	8	9	10	8
7	Combat Exmplry	8.74	1.83	63	70	58	55	54	50	.	72	24	19	8	-15	18	9	20	11	13	23	17	14	13	8	8	23
8	Combat Problems	10.72	1.95	59	60	58	49	50	46	72	.	21	16	7	-22	13	12	14	6	11	26	12	15	13	9	1	14
9	Awards & Certs	2.62	1.73	20	21	13	20	8	10	24	21	.	17	20	-4	9	-0	10	-1	-0	5	11	-0	-2	-2	5	1
10	Phys. Readiness	260.40	33.39	24	19	13	35	7	13	19	16	17	.	11	-9	5	1	6	5	0	-5	8	5	4	12	2	8
11	M16 Qualific.	1.86	0.80	4	2	4	5	-5	-1	8	7	20	11	.	3	2	-4	12	8	-6	7	3	-3	2	-7	-1	13
12	Articles 15	0.22	0.62	-23	-19	-27	-23	-21	-21	-15	-22	-4	-9	3	.	-42	-13	-5	1	-10	-7	2	-10	-5	-5	-5	4
13	Promotion Rate	0.01	0.46	20	19	19	20	21	17	18	13	9	5	2	-42	.	12	5	2	6	6	9	5	7	6	4	-0
14	HO Tech.	86.09	14.26	17	25	20	3	24	22	9	12	-0	1	-4	-13	12	.	28	13	58	34	33	58	25	23	7	11
15	HO Basic	18.56	5.00	14	14	10	7	8	13	20	14	10	6	12	-5	5	28	.	43	29	48	35	23	26	17	6	23
16	HO Safety	20.54	4.00	3	2	-3	-3	-2	1	11	6	-1	5	8	1	2	13	43	.	11	28	23	7	13	10	0	17
17	JK Tech.	42.21	9.53	22	29	22	1	28	22	13	11	-0	0	-6	-10	6	58	29	11	.	47	48	73	42	24	17	17
18	JK Basic	25.23	5.16	15	17	15	2	16	15	23	26	5	-5	7	-7	6	34	48	28	47	.	50	40	44	27	27	28
19	JK Safety	16.24	3.01	17	18	11	2	16	16	17	12	11	8	3	2	9	33	35	23	48	50	.	43	38	32	19	25
20	SK Tech.	44.99	9.78	21	28	20	-1	28	26	14	15	-0	5	-3	-10	5	58	23	7	73	40	43	.	44	33	15	16
21	SK Basic	9.90	2.28	13	17	7	0	10	8	13	13	-2	4	2	-5	7	25	26	13	42	44	38	44	.	32	18	31
22	SK Safety	4.26	1.29	11	9	8	-5	9	9	8	9	-2	12	-7	-5	6	23	17	10	24	27	32	33	32	.	4	15
23	SK Comm	0.38	0.48	5	7	1	0	6	10	8	1	5	2	-1	-5	4	7	6	0	17	27	19	15	18	4	.	11
24	SK Vehicle	2.71	1.21	10	11	4	-2	7	8	23	14	1	8	13	4	-0	11	23	17	17	28	25	16	31	15	11	.

N= 353

TABLE 8

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 91A: MEDICAL SPECIALIST

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	Overall Rating	4.61	0.82	.	86	78	60	67	62	71	70	22	15	-2	-29	32	17	6	13	28	24	25	4	8	28	12	15	6
2	Eff/Ldr Rating	4.40	0.77	86	.	76	56	73	67	73	71	24	13	-4	-30	33	20	9	19	26	25	21	-2	13	33	14	16	9
3	Discipline Rtng	4.54	0.91	78	76	.	47	60	47	58	69	12	7	-8	-29	31	15	11	13	28	21	20	-4	6	33	14	11	10
4	Fitness Rating	4.74	0.92	60	56	47	.	41	38	49	47	10	39	0	-20	18	3	-0	-0	3	7	4	7	1	-1	2	-4	-15
5	Job-Spec Tech	23.09	3.24	67	73	60	41	.	67	55	54	15	6	-1	-27	26	18	2	13	22	16	14	-3	3	32	5	15	7
6	Job-Spec Other	18.47	2.55	62	67	47	38	67	.	64	51	28	7	9	-17	27	10	6	16	18	25	20	5	15	23	11	18	16
7	Combat Exmpiry	9.20	1.48	71	73	58	49	55	64	.	79	30	9	9	-20	26	16	10	15	22	25	22	1	18	28	20	17	12
8	Combat Problems	10.11	1.77	70	71	69	47	54	51	79	.	23	5	-5	-28	30	14	6	11	24	22	23	-1	9	32	28	16	12
9	Awards & Certs	3.04	2.01	22	24	12	10	15	28	30	23	.	14	34	-6	13	3	7	22	4	10	8	11	18	4	11	12	6
10	Phys. Readiness	255.71	31.94	15	13	7	39	6	7	9	5	14	.	17	-11	-2	4	-6	-5	-3	-7	-2	3	-5	-8	-3	-6	-7
11	M16 Qualific.	2.08	0.78	-2	-4	-8	0	-1	9	9	-5	34	17	.	-1	-4	3	0	8	-8	5	-7	-0	12	-4	-2	2	2
12	Articles 15	0.41	0.89	-29	-30	-29	-20	-27	-17	-20	-28	-6	-11	-1	.	-33	-10	1	-7	-10	-7	-6	12	-5	-13	-16	-6	-1
13	Promotion Rate	-0.00	0.58	32	33	31	18	26	27	26	30	13	-2	-4	-33	.	10	9	7	16	20	9	-9	11	18	11	14	11
14	HO Tech.	50.48	10.02	17	20	15	3	18	10	16	14	3	4	3	-10	10	.	16	34	39	27	30	2	13	44	8	28	14
15	HO Basic	9.57	3.00	6	9	11	-0	2	6	10	6	7	-6	0	1	9	16	.	17	21	37	21	9	14	17	18	22	11
16	HO Safety	33.52	4.30	13	19	13	-0	13	16	15	11	22	-5	8	-7	7	34	17	.	32	32	33	3	17	30	10	33	18
17	JK Tech.	85.32	13.71	28	26	28	3	22	18	22	24	4	-3	-8	-10	16	39	21	32	.	54	76	13	16	67	20	48	22
18	JK Basic	15.19	3.63	24	25	21	7	16	25	25	22	10	-7	5	-7	20	27	37	32	54	.	55	8	24	41	23	38	22
19	JK Safety	42.71	7.35	25	21	20	4	14	20	22	23	8	-2	-7	-6	9	30	21	33	76	55	.	12	16	55	21	49	21
20	JK Vehicle	2.42	1.04	4	-2	-4	7	-3	5	1	-1	11	3	-0	12	-9	2	9	3	13	8	12	.	10	2	-3	6	6
21	JK Identify	6.62	2.32	8	13	6	1	3	15	18	9	18	-5	12	-5	11	13	14	17	16	24	16	10	.	15	15	13	13
22	SK Tech.	91.65	17.57	28	33	33	-1	32	23	29	32	4	-8	-4	-13	18	44	17	30	67	41	55	2	15	.	24	52	36
23	SK Basic	2.04	0.78	12	14	14	2	5	11	20	28	11	-3	-2	-16	11	8	18	10	20	23	21	-3	15	24	.	26	14
24	SK Safety	5.77	1.56	15	16	11	-4	15	18	17	16	12	-8	2	-6	14	28	22	33	48	38	49	6	13	52	26	.	27
25	SK Vehicle	4.51	1.62	6	9	10	-15	7	16	12	12	6	-7	2	-1	11	14	11	18	22	22	21	6	13	36	14	27	.

N= 372

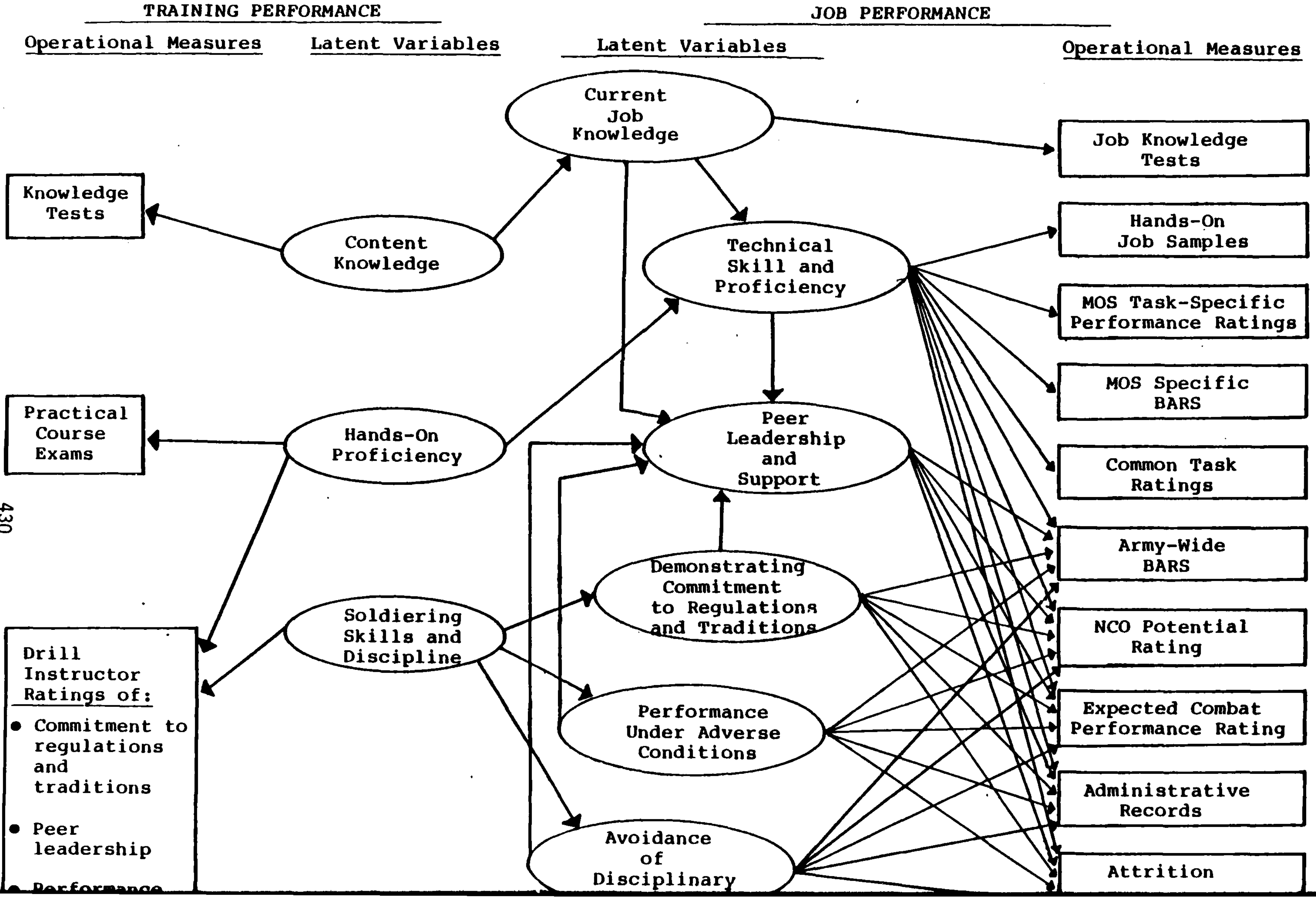
TABLE 9

JOB PERFORMANCE MEASURE SUMMARY STATISTICS
FOR 95B: MILITARY POLICE

#	VARIABLE	MN	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Overall Rating	4.74	0.80	.	87	69	70	78	68	74	70	18	22	13	-28	21	15	18	8	4	1	12	10	8	4	9	8	19	8	7	6
2	Eff/Ldr Rating	4.50	0.73	87	.	71	61	77	72	78	68	20	17	11	-22	19	14	21	10	10	1	10	13	7	7	15	10	18	13	12	8
3	Discipline Rtng	4.71	0.77	69	71	.	46	65	48	55	63	6	7	4	-27	26	6	9	5	7	-3	11	10	6	8	12	12	15	18	14	6
4	Fitness Rating	4.90	0.84	70	61	46	.	58	56	55	52	16	43	13	-26	16	9	12	7	5	2	0	3	-2	-0	3	-3	3	5	2	-1
5	Job-Spec Tech	29.00	3.66	78	77	65	58	.	73	68	63	15	15	11	-19	17	12	19	8	12	2	14	16	9	5	9	7	17	11	12	5
6	Job-Spec Other	23.60	3.10	68	72	48	56	73	.	71	61	32	18	27	-16	8	11	22	14	19	7	2	13	6	4	16	7	9	12	6	5
7	Combat Exmpiry	9.56	1.36	74	78	55	55	68	71	.	79	19	19	16	-17	14	17	19	14	14	6	10	17	11	7	15	6	19	15	9	9
8	Combat Problems	10.45	1.53	70	68	63	52	63	61	79	.	15	15	15	-28	21	14	16	11	10	-0	16	18	17	11	14	10	19	15	9	6
9	Awards & Certs	3.17	2.09	18	20	6	16	15	32	19	15	.	20	26	-3	11	8	16	6	8	11	-11	2	-1	7	-0	9	4	4	10	4
10	Phys. Readiness	251.75	32.78	22	17	7	43	15	18	19	15	20	.	13	-12	7	-1	6	4	2	4	-6	1	-3	-3	-2	-12	-2	-8	-4	-3
11	M16 Qualific.	2.28	0.76	13	11	4	13	11	27	16	15	26	13	.	1	-3	4	6	5	4	6	-3	7	3	2	-9	-1	1	-0	-2	-2
12	Articles 15	0.27	0.70	-28	-22	-27	-26	-19	-16	-17	-28	-3	-12	1	.	-39	-4	-0	-8	-3	2	-8	-8	-5	-2	0	0	-7	-6	-3	5
13	Promotion Rate	0.01	0.47	21	19	26	16	17	8	14	21	11	7	-3	-39	.	4	4	6	1	-3	15	15	16	10	6	2	0	10	7	1
14	HO Tech.	31.58	4.63	15	14	6	9	12	11	17	14	8	-1	4	-4	4	.	18	12	6	11	13	11	10	7	3	5	14	7	3	10
15	HO Basic	50.04	10.28	18	21	9	12	19	22	19	16	16	6	6	-0	4	18	.	20	21	18	18	34	26	21	17	5	12	27	23	12
16	HO Safety	31.76	5.16	8	10	5	7	8	14	14	11	6	4	5	-8	6	12	20	.	9	15	10	20	21	21	12	9	15	17	19	9
17	HO Comm	10.57	2.17	4	10	7	5	12	19	14	10	8	2	4	-3	1	6	21	9	.	31	14	21	13	30	14	7	9	21	16	17
18	HO Vehicle	10.56	1.63	1	1	-3	2	2	7	6	-0	11	4	6	2	-3	11	18	15	31	.	1	4	8	19	16	2	11	12	9	13
19	JK Tech.	38.44	5.90	12	10	11	0	14	2	10	16	-11	-6	-3	-8	15	13	18	10	14	1	.	60	53	35	18	15	40	33	28	24
20	JK Basic	50.11	9.99	10	13	10	3	16	13	17	18	2	1	7	-8	15	11	34	20	21	4	60	.	60	51	32	22	38	49	46	35
21	JK Safety	25.52	4.55	8	7	6	-2	9	6	11	17	-1	-3	3	-5	16	10	26	21	13	8	53	60	.	40	24	20	36	37	33	27
22	JK Comm	13.54	4.62	4	7	8	-0	5	4	7	11	7	-3	2	-2	10	7	21	21	30	19	35	51	40	.	26	18	22	33	31	36
23	JK Vehicle	2.03	1.19	9	15	12	3	9	16	15	14	-0	-2	-9	0	6	3	17	12	14	16	18	32	24	26	.	15	18	23	23	20
24	JK Identify	6.88	2.29	8	10	12	-3	7	7	6	10	9	-12	-1	0	2	5	5	9	7	2	15	22	20	18	15	.	21	20	21	17
25	SK Tech.	40.20	7.04	19	18	15	3	17	9	19	19	4	-2	1	-7	0	14	12	15	9	11	40	38	36	22	18	21	.	49	49	38
26	SK Basic	17.85	3.66	9	13	18	5	11	12	15	15	4	-8	-0	-6	10	7	27	17	21	12	33	49	37	33	23	20	49	.	60	40
27	SK Safety	14.45	3.35	7	12	14	2	12	6	9	9	10	-4	-2	-3	7	3	23	18	16	9	28	46	38	31	23	21	49	60	.	39
28	SK Comm	3.12	1.23	6	8	6	-1	5	5	9	6	4	-3	-2	5	1	10	12	9	17	13	24	35	27	36	20	17	33	40	39	.
29	SK Vehicle	6.02	1.90	8	9	10	1	5	4	8	13	9	2	-1	-7	2	6	15	10	11	10	19	31	28	24	17	12	37	44	40	32
30	SK Identify	0.29	0.51	-6	-5	-3	-7	-2	-2	1	0	-0	-5	4	6	-1	-0	-7	-6	-10	-1	1	4	0	-3	-4	-2	2	-1	-6	-0

N= 506

Figure 2


TRAINING PERFORMANCE
Operational Measures
Latent Variables
JOB PERFORMANCE
Latent Variables
Operational Measures
Knowledge Tests
Practical Course Exams
Drill Instructor Ratings of:
• Commitment to regulations and traditions
• Peer leadership
• Performance
Content Knowledge
Hands-On Proficiency
Soldiering Skills and Discipline
Current Job Knowledge
Technical Skill and Proficiency
Peer Leadership and Support
Demonstrating Commitment to Regulations and Traditions
Performance Under Adverse Conditions
Avoidance of Disciplinary
Job Knowledge Tests
Hands-On Job Samples
MOS Task-Specific Performance Ratings
MOS Specific BARS
Common Task Ratings
Army-Wide BARS
NCO Potential Rating
Expected Combat Performance Rating
Administrative Records
Attrition

fewer different scales for this method. The emergence of method factors was fully anticipated and was consistent with prior findings (e.g., Landy & Farr, 1980).

The second consistent result was a rationally satisfying correspondence between the administrative measures scales and the three Army-wide rating factors. The awards and certificates scale from the administrative measures loaded together with the Army-wide effort/leadership rating factor; the Article 15 and promotion rate scale loaded with the personal discipline factor (most of the variance in promotion rate was thought to be due to retarded advancement associated with disciplinary problems); and the physical readiness scale loaded with the fitness/appearance factor.

The third consistent result was that the ratings of job performance rarely loaded together with the hands-on and written measures. This finding, coupled with the correspondence between the rating scales and the administrative measures noted above, suggested that the rating scales were tapping the motivation and effort components of job performance (i.e., typical performance), while the hands-on and written measures were assessing job knowledge and skill (i.e., maximal performance).

The final observation from the empirical factor analyses was that, with the possible exception of the technical skills factor, there was not much evidence that the six content category factors crossed measurement methods. The hands-on communication score, for example, was likely to be as correlated with the written safety score as with the written communication score. This result was taken as evidence against separate content categories of job knowledge and skill within the common task domain.

Based on these findings from the empirical analyses, a revised model was constructed to account for the correlations among our performance measures. This model included five job performance constructs and two measurement method factors. These were:

- job-specific technical knowledge and skill
- general soldiering knowledge and skill
- effort and leadership
- personal discipline
- physical fitness and appearance
- written test "method" factor

- ratings "method" factor

Several minor issues remained before we could test the model for goodness of fit within our nine jobs. One was whether the job-specific ratings scales were measuring job-specific technical knowledge and skill, or effort and leadership, or both. The intercorrelations among our performance factors suggested that these rating scales were measuring both of these performance constructs, though they seemed to correlate more highly with other measures of effort and leadership than with measures of job-specific technical knowledge and skill.

A second issue was whether it was necessary to posit hands-on and administrative measures "method" factors in order to account for the intercorrelations within each of these sets of measures. The average intercorrelation among the scores within each of these sets was not particularly high. Therefore, for the sake of parsimony, we decided to try to fit a model without the two additional factors.

Finally, a third issue was how to incorporate the M16 qualification score from the administrative measures into our model. The score did not correlate consistently with any of the other performance measures, and its psychometric properties were somewhat questionable. However, it seemed to measure an important aspect of first-term soldier performance. Therefore, we created a sixth job performance construct, M16 qualification. The M16 score was the only measures assigned to this construct.

CONFIRMING THE MODEL WITHIN EACH JOB

The next step in the analysis was to conduct separate tests of goodness of fit of this target model within each of the nine jobs. This was done using the LISREL confirmatory factor analysis program program (Joreskog & Sorbom, 1981).

In conducting a confirmatory factor analysis with LISREL, it is necessary to specify the structure of three different parameter matrices: Lambda-Y, the hypothesized factor structure matrix (a matrix of regression coefficients for predicting the observed variables from the underlying latent constructs); Theta-Epsilon, the matrix of uniqueness or error components (and intercorrelations); and Psi, the matrix of covariances among the factors. In these analyses, we set the diagonal elements of Psi (i.e., the factor variances) to one, forcing a "standardized" solution. This meant that the off-diagonal elements in Psi would represent the correlations among and between our performance constructs and method factors. We further specified that the correlation among the two method factors and each performance construct should be zero. This effectively defined the method factor as that portion of the common variance among measures from the same method that was not predictable from (i.e., correlated with) any of the other related factor or performance construct scores.

A few technical problems were encountered in fitting the hypothesized model for several of the jobs. Solutions were obtained with some factor loadings greater than one and with negative uniqueness estimates for the corresponding observed variables. In addition, estimates of the correlations among the performance constructs occasionally exceeded unity. These problems were resolved by computing the squared multiple correlation (SMC) for predicting each observed variable from all of the other variables, and setting the uniqueness estimates (i.e., Theta-Epsilon diagonal) to one minus this SMC. This approach eliminated all factor loadings and correlations greater than one. In most cases, a second "iteration" was performed to adjust the initial Theta-Epsilon estimates so that the diagonal of the estimated correlation matrix would be as close to one as possible.

Table 10 shows the final factor loading estimates from Lambda-Y for each job. Tables 11 and 12 show the uniqueness estimates from Theta-Epsilon and the factor intercorrelation estimates from Psi, respectively.

LISREL also computes a goodness-of-fit index based on a comparison of the actual correlations among the observed variables and the correlations estimated from Lambda-Y, Theta-Epsilon, and Psi. The goodness of fit is distributed as chi-square, with degrees of freedom dependent on the number of observed variables and the number of parameters estimated. The expected value of chi-square is equal to the degrees of freedom. If chi-square is significantly greater than the degrees of freedom, it is a sign that the model does not fit the correlations among the observed variables.

Table 13 shows the value of chi-square for each job. These chi-square values should be interpreted with considerable caution. The approach we used was not purely confirmatory. The hypothesized target model was based in part on analyses of these same data. In addition, LISREL was "told" that the Theta-Epsilon (uniqueness) parameters were all fixed, and therefore did not "use up" any degrees of freedom estimating these parameters; in fact, these values were estimated entirely from the data.

Table 10

FACTOR LOADINGS
(Separate Model for Each Job)

	Military Occupational Specialty								
Factor/Variable	11B	13B	19E	31C	63B	64C	71L	91A	95B
Tech. Job Knowledge									
HO Tech	--	.61	.47	.64	.51	.29	.77	.59	.32
JK Tech	--	.75	.78	.79	.74	.26	.78	.75	.32
SK Tech	--	.70	.79	.73	.82	.55	.29	.81	.43
MOS Tech Rtngs	--	.45	.10	.22	.25	.25	.34	.10	.13
General Soldiering									
HO Basic	.60	.51	.46	.64	.17	.50	.60	.42	.60
HO Safety	.26	.33	.32	.31	.12	.63	.37	.48	.47
HO Comm	.05	.06	.39	.56	--	--	--	--	.80
HO Vehicle	--	--	--	.22	.17	**	--	--	.31
JK Basic	.76	.52	.74	.62	.45	.48	.87	.58	.46
JK Safety	.55	.37	.75	.38	.71	.51	.72	.58	.33
JK Comm	.30	.23	.66	.38	--	--	--	--	.29
JK Vehicle	--	.17	--	.10	.41	**	--	--	.35
JK Identify	.46	--	.20	.28	--	.12	--	.24	.21
SK Basic	.73	.45	.67	.39	.78	.56	.45	.44	.42
SK Safety	.47	.32	.53	.62	.57	.47	.30	.64	.32
SK Comm	.42	.26	.42	--	.41	.35	.20	--	.20
SK Vehicle	.22	.24	.05	.30	.61	**	.22	.47	.28
SK Identify	.46	--	.46	.13	--	--	--	--	--
Effort/Leadership									
Eff/Ldr Rating	.76	.56	.85	.64	.68	.83	.66	.76	.70
MOS Tech Rtngs	.70	--	.63	.40	.41	.50	.25	.59	.52
MOS Other Rtngs	.77	.41	.48	.43	.54	.62	.43	.61	.56
Comb. Exmplry	.80	.47	.68	.54	.57	.87	.63	.80	.77
Comb. Problems	.48	.20	--	.39	.52	.53	.55	--	.56
Awards/Cert.	.32	.23	.24	.19	.28	.25	.34	.34	.22
Overall Rating	.46	.39	.33	.17	.57	.42	.65	--	.41

Table 10 (Continued)

FACTOR LOADINGS
(Separate Model for Each Job)

	Military Occupational Specialty								
Factor/Variable	11B	13B	19E	31C	63B	64C	71L	91A	95B
Discipline									
Discpln. Rtngs	.77	.58	.73	.45	.63	.85	.74	.58	.73
Comb. Problems	.29	.16	.62	.03	.05	.19	--	.82	.33
Articles 15	-.63	-.61	-.55	-.62	-.65	-.47	-.69	-.46	-.60
Promotion Rate	.74	.61	.68	.79	.63	.57	.59	.54	.54
Overall Rating	.39	.20	.53	.54	.09	.42	.06	.75	.38
Fitness									
Fitness Rtngs	.69	.23	.84	.48	.54	.42	.50	.60	.78
Phys. Readiness	.11	.90	.49	.89	.70	.53	.76	.69	.69
Ratings Meth.									
AW Rtngs	.60	.73	.47	.70	.66	.54	.65	.66	.66
MOS Rtngs	.73	.73	.60	.69	.67	.49	.69	.54	.63
Comb. Rtngs	.47	.65	.55	.69	.57	.27	.55	.47	.40
Written Meth.									
JK Tech	--	.47	.28	.55	.59	.73	.44	.58	.57
JK Basic	.41	.51	.33	.40	.61	.57	.11	.37	59
JK Safety	.37	.52	.12	.63	.08	.49	.17	.76	.57
JK Comm.	.34	.11	.07	.55	--	--	--	--	.52
JK Vehicle	--	--	--	.42	.62	**	--	.24	.21
JK Identify	-.15	.23	.50	.36	--	.05	--	.08	.23
SK Tech.	--	.48	.48	.55	.46	.88	.42	.27	.50
SK Basic	.50	.66	.54	.59	.15	.51	.54	--	.60
SK Safety	.53	.55	.42	.29	.34	.48	.44	.19	.60
SK Comm.	.51	.47	.46	--	.16	.24	.05	--	.54
SK Vehicle	.49	.57	.24	.48	.55	**	.38	.05	.42
SK Identify	.21	--	.42	.44	--	--	--	--	--
M16	.71	.71	.71	.71	.71	.71	.71	.71	

** - vehicle content was merged into MOS technical for 64C.

Table 11

UNIQUENESS ESTIMATES
(Separate Model for Each Job)

Observed Variable	Military Occupational Specialty								
	11B	13B	19E	31C	63B	64C	71L	91A	95B
HO Tech	--	.52	.71	.48	.64	.74	.33	.57	.88
HO Basic	.59	.66	.75	.52	.95	.74	.55	.76	.63
HO Safety	.92	.85	.75	.52	.95	.59	.79	.71	.77
HO Comm	.95	.95	.81	.62	--	--	--	--	.82
HO Vehicle	--	--	--	.03	.95	**	--	--	.90
JK Tech	--	.21	.30	.15	.12	.39	.17	.11	.53
JK Basic	.10	.43	.22	.26	.29	.44	.31	.58	.43
JK Safety	.32	.53	.32	.31	.45	.49	.44	.15	.57
JK Comm.	.56	.93	.32	.34	--	--	--	--	.64
JK Vehicle	--	--	--	.56	.32	**	--	.94	.82
JK Identify	.36	.89	.40	.51	--	.95	--	.92	.90
SK Tech.	--	.27	.13	.09	.10	.14	.14	.15	.52
SK Basic	.09	.37	.14	.48	.31	.42	.54	.74	.46
SK Safety	.46	.59	.43	.41	.50	.55	.72	.47	.55
SK Comm.	.40	.72	.35	--	.65	.82	.78	--	.67
SK Vehicle	.73	.62	.69	.55	.18	**	.73	.76	.75
SK Identify	.69	--	.42	.68	--	--	--	--	--
Overall Rating	.13	.13	.13	.13	.13	.13	.13	.13	.18
Eff/Ldr Rating	.11	.11	.11	.11	.11	.05	.11	.11	.05
Discpln. Rtngs	.22	.22	.22	.22	.22	.05	.22	.22	.06
Fitness Rtngs	.38	.38	.38	.38	.38	.05	.38	.38	.05
MOS Tech Rtngs	.08	.11	.13	.14	.08	.37	.17	.12	.33
MOS Other Rtngs	.10	.13	.17	.19	.12	.35	.20	.18	.27
Comb. Exmplry	.02	.02	.02	.02	.02	.14	.02	.02	.08
Comb. Problems	.13	.13	.13	.13	.13	.60	.13	.13	.40
Awards/Cert.	.89	.94	.93	.95	.91	.94	.86	.85	.90
Phys. Readiness	.95	.33	.67	.34	.50	.83	.46	.49	.49
Articles 15	.58	.59	.68	.60	.56	.76	.51	.75	.64
Promotion Rate	.45	.60	.53	.41	.57	.64	.62	.67	.70
M16	.50	.50	.50	.50	.50	.50	.50	.50	.00

** - vehicle content was merged into MOS technical for 64C.

Table 12

ESTIMATED FACTOR CORRELATIONS
(Separate Model for Each Job)

		Military Occupational Specialty								
1st Factor	2nd Factor	11B	13B	19E	31C	63B	64C	71L	91A	95B
Tech. Job. Knowledge	General Soldiering	n/a	.77	.83	.63	.58	.73	.48	.66	.70
	Effort/ Leadership	n/a	.86	.51	.44	.50	.78	.44	.35	.46
	Discipline	n/a	.13	.37	.26	.12	.69	.19	.43	.50
	Fitness	n/a	.01	.03	.04	-.18	-.09	.10	-.05	-.09
	M16	n/a	.00	.04	.11	.05	.05	-.09	-.17	-.10
General Soldiering	Effort/ Leadership	.67	.89	.58	.57	.53	.44	.37	.43	.40
	Discipline	.42	.29	.45	.30	.29	.29	.04	.37	.24
	Fitness	.25	-.19	.05	-.05	-.03	-.14	.09	-.05	.00
	M16	.27	-.06	.30	.30	.04	.11	.27	.02	.02
Effort/ Leadership	Discipline	.49	.67	.62	.55	.65	.51	.51	.59	.39
	Fitness	.57	.04	.38	-.11	.10	.23	.32	.21	.42
	M16	.38	-.13	.21	.24	-.02	.35	.22	.17	.28
Discipline	Fitness	.33	.05	.24	.24	.30	.30	. 7	.19	.2
	M16	-.12	-.25	-.30	.09	-.28	-.11	.01	-.28	-.0
Fitness	M16	.52	.26	-.05	.02	.19	.22	.18	.27	.2

Table 13

GOODNESS-OF-FIT INDICES
(Separate Model for Each Job)

JOB	Root Mean Square Residual	Chi-Square	DF	P
11B: Infantryman	.061	326.2	227	.02
13B: Cannon Crewman	.057	350.0	322	.14
19E: Tank Crewman	.065	170.0	348	.999
31C: Radio/Teletype Operator	.069	369.2	375	.58
63B: Vehicle/Generator Mechanic	.060	332.1	296	.07
64C: Motor Transport Operator	.058	280.1	247	.07
71L: Administrative Clerk	.067	232.6	249	.77
91A: Medical Corpsman	.061	277.1	275	.45
95B: Military Policeman	.052	470.0	374	.001

CONFIRMATION OF THE OVERALL MODEL

The results of the confirmatory procedures applied to the performance measures from each job generally supported a common structure of job performance. The procedures also yielded reasonably similar estimates of the intercorrelations among the constructs and of the loadings of the observed variables on these constructs across the nine jobs.

The final step in our analyses was to determine whether the variation in some of these parameters across jobs could be attributed to sampling variation. The specific model that we explored stated that (1) the correlation among factors was invariant across jobs and (2) the loadings of all of the Army-wide measures on the performance constructs and on the rating method factor were also constant across jobs. Since different items and scales were used for different jobs in the job-specific measures, it was not reasonable to expect constant measurement precision (i.e., the same balance between factor loadings and uniquenesses) across jobs.

The overall model tested was relatively strong. It was quite possible that selectivity differences in the different jobs would lead to differences in the apparent measurement precision of the common instruments or differences in the correlations between the constructs. This would tend to make it appear that the different jobs required different performance models, when in fact they do not.

The LISREL multi-groups option was used to test the overall model. This option requires that the number of observed variables be the same for each job. This was a problem, since virtually every job was missing scores on at least one of the six construct categories for at least one of the three knowledge and skill measurement methods. To handle this problem, the Theta-Epsilon error estimates for these variables were set to 1.00, and the observed correlations between these variables and all the other variables were set to zero. It was thus necessary to count the number of "observed" correlations that we generated in this manner and subtract this number from the degrees of freedom when determining the significance of the chi-square goodness-of-fit statistic.

The overall model fit extremely well. The root mean square residual was .047, and the chi-square was 2508.1. There were 2403 degrees of freedom after adjusting for missing variables and the use of the data in estimating uniquenesses. This yields a significance level of .07, not enough to reject the model. Table 14 shows the estimated intercorrelations between the construct scores. As noted

Table 14

ESTIMATED FACTOR CORRELATIONS
(Overall Model)

First Factor	Second Factor	Correlation
Technical Job Knowledge	General Soldiering	.80
	Effort/Leadership	.48
	Discipline	.35
	Fitness	.01
General Soldering	Effort/Leadership	.47
	Discipline	.35
	Fitness	.06
Effort/Leadership	Discipline	.67
	Fitness	.42
Discipline	Fitness	.40

previously, these were constrained to be equal for all nine jobs. Tables 15 and 16 show the factor loadings and uniquenesses for each job under this constrained model.

SUMMARY AND DISCUSSION

A set of up to 29 performance measures was identified from a wide battery of performance measures representing different measurement methods and different aspects of job performance. A confirmatory approach was used to determine the extent to which dimensions of individual variation in these job performance measures could be explained by a common set of performance factors for nine different jobs. The results indicated that they could.

Several aspects of the final structure are noteworthy. First, in spite of the confounding of some of the performance with measurement method, the latent performance structure appears to be composed of very distinct components. It is reasonable to expect that the different performance constructs would be predicted by different things, so that validity generalization may not exist across the performance constructs within a job. If this is so, there is a genuine question of how the performance constructs should be weighted in forming an overall appraisal of performance for use in personnel decisions.

It is tempting to infer that Effort/Leadership and Maintaining Personal Discipline, particularly the latter, reflect aspects of performance that are under motivational control and consequently may be better predicted by personality or interest measures than by measures of ability or skill. This leads us to the question of whether choices such as showing up on time, staying out of trouble, and expending extra effort under adverse conditions are a function of state or trait variables. We do have considerable data to focus on the question. It will be particularly interesting to examine the ability and personality correlates of the method factors as compared to the correlates of the content factors and of the content factors that have had the method variance partialled out. For example, the nature of ratings "halo" may become a bit clearer.

Finally, since (a) the five-factor solution is stable across jobs sampled from this population, (b) the performance constructs seek to make sense, and (c) the constructs are based on measures carefully developed to be content valid, it seems safe to ascribe some degree of construct validity to them.

Table 15

FACTOR LOADINGS
(Overall Model)

	Military Occupational Specialty								
Factor/Variable	11B	13B	19E	31C	63B	64C	71L	91A	95B
Tech. Job Knowledge									
HO Tech	n/a	.59	.43	.58	.46	.27	.71	.54	.29
JK Tech	n/a	.71	.79	.76	.57	.72	.70	.74	.37
SK Tech	n/a	.66	.70	.54	.73	.55	.68	.85	.42
MOS Tech Rtngs	n/a	.21	.12	.16	.25	.01	.12	.05	-.02
General Soldiering									
HO Basic	.52	.66	.44	.52	.16	.51	.57	.35	.58
HO Safety	.20	.44	.31	.36	.10	.49	.30	.50	.41
HO Comm	.06	.12	.37	.52	n/a	n/a	n/a	n/a	.43
HO Vehicle	n/a	n/a	n/a	.15	.21	**	n/a	n/a	.27
JK Basic	.95	.50	.79	.64	.42	.69	.66	.69	.49
JK Safety	.69	.36	.75	.45	.53	.66	.57	.65	.42
JK Comm	.35	.25	.59	.51	n/a	n/a	n/a	n/a	.39
JK Vehicle	n/a	n/a	n/a	.28	.37	**	n/a	.07	.34
JK Identify	.43	.21	.34	.36	n/a	.12	n/a	.39	.18
SK Basic	.81	.40	.67	.33	.70	.50	.42	.40	.38
SK Safety	.57	.34	.45	.40	.63	.43	.31	.62	.34
SK Comm	.51	.21	.31	n/a	.42	.29	.17	n/a	.23
SK Vehicle	.35	.22	.06	.17	.65	**	.32	.36	.21
Effort/Leadership									
Eff/Ldr Rating*	.76	.76	.76	.76	.76	.76	.76	.76	.76
MOS Tech Rtngs	.59	.33	.54	.50	.45	.62	.43	.62	.62
MOS Other Rtngs	.77	.59	.33	.45	.59	.48	.47	.58	.58
Comb. Exmplry*	.72	.72	.72	.72	.72	.72	.72	.72	.72
Comb. Problems*	.44	.44	.44	.44	.44	.44	.44	.44	.44
Awards/Cert.*	.26	.26	.26	.26	.26.	.26	.26	.26	.26
Overall Rating*	.48	.48	.48	.48	.48	.48	.48	.48	.48
Discipline									
Discpln. Rtngs*	.69	.69	.69	.69	.69	.69	.69	.69	.69
Comb. Problems*	.25	.25	.25	.25	.25	.25	.25	.25	.25
Articles 15*	-.48	-.48	-.48	-.48	-.48	-.48	-.48	-.48	-.48
Promotion Rate*	.52	.52	.52	.52	.52	.52	.52	.52	.52
Overall Rating*	.28	.28	.28	.28	.28	.28	.28	.28	.28

Table 15 (Continued)

FACTOR LOADINGS
(Overall Model)

	Military Occupational Specialty								
Factor/Variable	11B	13B	19E	31C	63B	64C	71L	91A	95B
Fitness									
Fitness Rtngs*	.82	.82	.82	.82	.82	.82	.82	.82	.82
Phys. Readiness*	.37	.37	.37	.37	.37	.37	.37	.37	.37
Ratings Meth.									
AW Rtngs*	.56	.56	.56	.56	.56	.56	.56	.56	.56
MOS Rtngs*	.61	.61	.61	.61	.61	.61	.61	.61	.61
Comb. Rtngs*	.42	.42	.42	.42	.42	.42	.42	.42	.42
Written Meth.									
JK Tech	n/a	.49	.29	.54	.71	.30	.42	.49	.49
JK Basic	-.16	.51	.29	.40	.53	.25	.28	.60	.60
JK Safety	-.07	.49	.07	.52	.26	.28	.35	.52	.52
JK Comm.	.00	.11	.19	.38	n/a	n/a	n/a	.41	.41
JK Vehicle	n/a	n/a	n/a	.19	.62	**	n/a	.20	.20
JK Identify	-.05	.20	.12	.17	n/a	.10	n/a	.25	.25
SK Tech.	n/a	.54	.65	.64	.49	.71	.45	.53	.53
SK Basic	.44	.68	.58	.61	.25	.66	.50	.60	.60
SK Safety	.34	.51	.49	.57	.18	.56	.30	.59	.59
SK Comm.	.51	.46	.60	n/a	.20	.36	.20	.50	.50
SK Vehicle	.38	.51	.17	.60	.45	**	.17	.46	.46

* - constrained equal across MOS.
** - vehicle content was merged into MOS technical for 64C.

Table 16

UNIQUENESS ESTIMATES
(Overall Model)

Observed	Military Occupational Specialty								
Variable	11B	13B	19E	31C	63B	64C	71L	91A	95B
HO Tech	n/a	.62	.79	.62	.76	.91	.44	.68	.90
HO Basic	.72	.58	.80	.70	.95	.73	.64	.87	.67
HO Safety	.95	.84	.90	.87	.95	.73	.90	.75	.81
HO Comm	.95	.95	.86	.71	n/a	n/a	n/a	n/a	.82
HO Vehicle	n/a	n/a	n/a	.95	.95	**	n/a	n/a	.93
JK Tech	n/a	.23	.28	.13	.15	.32	.28	.16	.60
JK Basic	.10	.44	.28	.40	.48	.41	.44	.47	.40
JK Safety	.48	.56	.41	.49	.62	.44	.55	.26	.54
JK Comm.	.85	.91	.57	.55	n/a	n/a	n/a	n/a	.67
JK Vehicle	n/a	n/a	n/a	.87	.44	**	n/a	.95	.85
JK Identify	.71	.90	.84	.81	n/a	.95	n/a	.64	.90
SK Tech.	n/a	.25	.10	.24	.18	.17	.27	.19	.54
SK Basic	.13	.37	.20	.52	.41	.31	.58	.83	.49
SK Safety	.54	.62	.54	.51	.55	.51	.80	.29	.54
SK Comm.	.46	.75	.48	n/a	.77	.78	.92	n/a	.70
SK Vehicle	.75	.68	.95	.61	.31	**	.86	.86	.75
Overall Rating*	.18	.18	.18	.18	.18	.18	.18	.18	.18
Eff/Ldr Rating*	.09	.09	.09	.09	.09	.09	.09	.09	.09
Discpln. Rtngs*	.17	.17	.17	.17	.17	.17	.17	.17	.17
Fitness Rtngs*	.05	.05	.05	.05	.05	.05	.05	.05	.05
MOS Tech Rtngs	.18	.34	.22	.24	.18	.18	.18	.18	.25
MOS Other Rtngs	.05	.24	.46	.37	.05	.05	.05	.05	.27
Comb. Exmplry*	.26	.26	.26	.26	.26	.26	.26	.26	.26
Comb. Problems*	.29	.29	.29	.29	.29	.29	.29	.29	.29
Awards/Cert.*	.93	.93	.93	.93	.93	.93	.93	.93	.93
Phys. Readiness*	.83	.83	.83	.83	.83	.83	.83	.83	.83
Articles 15*	.77	.77	.77	.77	.77	.77	.77	.77	.77
Promotion Rate*	.70	.70	.70	.70	.70	.70	.70	.70	.70

* - constrained equal across MOS.
** - vehicle content was merged into MOS technical for 64C.

Given the high degree of consistency across jobs in the structure of the performance measures, it is worth asking to what extent our performance model generalizes to even wider domains of jobs. Some limitations appear likely. The "general soldiering skills" constructs would almost surely be quite different outside the military. Perhaps it would be replaced by a more generalized job skill construct. Similarly, it is likely that the physical fitness and military appearance construct also would be somewhat different for civilian occupations. The remaining constructs --technical skill, effort and leadership, and personal discipline--all appear to be basic components of almost any job.

In generalizing to a wider domain of jobs, it is reasonable to suppose that other latent structures would fit other "populations" of jobs. For example, jobs that are not organized into units and that involve a great deal of written or oral communication (e.g., sales jobs) might have a different structure. It is tempting to ask how many different performance dimension structures define different populations of jobs. Such questions go well beyond the present finding, however, which is that a single structure did fit the jobs studied.

References:

Campbell, J.P., & Harris, J.H. (1985). Criterion reduction and combination via a particcipation decision-making panel. Paper presented at the 93rd Annual Meeting of the American Psychological Association, Los Angeles.

Joreskog, K.C., & Sorbom, D. (1981). LISREL VI: Analysis of Linear squares methods. Uppsala, Sweden: University of Uppsala.

Landy, F.J., & Farr, J.L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Working Paper

ARMY RESEARCH TO LINK STANDARDS FOR ENLISTMENT TO ON-THE-JOB PERFORMANCE (FIFTH ANNUAL REPORT TO CONGRESS)

Jane Arabian, Michael Rumsey
SELECTION AND CLASSIFICATION TECHNICAL AREA

and

Jeff McHenry
AMERICAN INSTITUTES FOR RESEARCH

September 1986

Reviewed by
Frances Grafton
Selection and Classification
Technical Area

Approved By
Lawrence Hanser, Chief
Selection and Classification
Technical Area

U.S. Army Research Institute
for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria VA 22333

This working paper is an unofficial document intended for limited distribution to obtain comments. The views, opinions, and/or findings contained in this document are those of the author(s) and should not be construed as the *official position of* ARI *or as* an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

CHAPTER 5

ARMY RESEARCH TO LINK STANDARDS FOR ENLISTMENT TO ON-THE-JOB PERFORMANCE

ARMY RESEARCH OBJECTIVES

Overall Army Goals

The Army Research Institute is currently engaged in a large-scale, multi-year project to improve the Army selection and classification system and, thereby, increase the overall effectiveness of the force. The goal of the Army's program for increasing the efficiency of enlisted personnel selection and utilization is to enable the Army to meet its peacetime and mobilization missions through improved matching of individuals to Military Occupational Specialties (MOS). The research is aimed at developing comprehensive selection and classification procedures to predict validly performance in Army training and occupational specialties. Specifically, this project will:

1. Validate existing selection measures against both existing and project-developed criteria. The criteria will include both Army-wide performance measures based on newly developed rating scales and direct measures of MOS-specific task performance.

2. Develop and validate new and/or improved selection and classification measures.

3. Validate intermediate criteria, such as performance in training, as predictors of subsequent criteria, e.g., job performance ratings, so that informed reassignment and promotion decisions can be made throughout an individual's tour.

4. Determine the relative utility to the Army of different performance levels across MOS.

5. Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

A complete description of the Army's research program and accomplishments have been published separately by the U.S. Army

Research Institute for the Behavioral and Social Sciences in the annual reports for Project A, "Improving the Selection, Classification and Utilization of Army Enlisted Personnel" (1983, 1984, 1985). A detailed description of the project also appears in previous Annual Reports to Congress on Joint-Service Efforts to Link Standards of Enlistment to On-the-Job Performance (1983, 1984, 1985).

Joint Project Goals

The Army goals and the joint project goals are the same.

Military Occupational Specialties Selected for Joint Study

The Army's research focuses on 19 MOS. The MOS were selected to be representative of the Army and include all operational ASVAB aptitude area composites. Blacks, Whites, Hispanics, males, and females are present in these MOS in the same proportions as in total accessions. These MOS represent 44% of annual Army enlistments.

A number of performance measures, including measures of training success, service-wide performance, and MOS-specific hands-on performance, were developed for these MOS. For reasons of cost efficiency, not all measures were developed for all 19 MOS. All project criterion measures were developed for the following MOS and serve as the focus of this report:

1. 11B - Infantryman
2. 13B - Cannon Crewman
3. 19E - Tank Crewman
4. 31C - Radio Teletype Operator
5. 63B - Light Wheel Vehicle/Power Generation Mechanic
6. 64C - Motor Transport Operator
7. 71L - Administrative Specialist
8. 91A - Medical Specialist
9. 95B - Military Police

Measures of training success and service-wide performance were developed for the following specialties:

1. 12B - Combat Engineer
2. 16S - MANPADS Crewman
3. 27E - TOW/Dragon Repairman
4. 51B - Carpentry/Masonry Specialist
5. 54E - Chemical Operations Specialist
6. 55B - Ammunition Specialist
7. 67N - Utility Helicopter Repairman
8. 76W - Petroleum Supply Specialist
9. 76Y - Supply Specialist
10. 94B - Food Service Specialist

RESEARCH DESIGN

Types of Measures to be Developed and Descriptions

Hands-on performance measures, job knowledge tests, and performance rating scales were developed for training success, service-wide performance, and MOS-specific performance for the MOS listed above. The Army's rationale for the development of multiple measures of job performance is based upon the knowledge that a soldier's job is multi-faceted and there are different aspects of job performance. Therefore, the Army's research project has developed different kinds of tests to assess these different aspects of job performance and thereby obtain information about the domain of job performance behaviors. A more complete description of these measures appears in the Project A Annual Reports (1983, 1984, 1985) and in the previous Annual Reports to Congress (1983, 1984, 1985).

Measures Construction. Construction of the different measures was based on job task analysis data. A multi-method job analysis approach was employed in which requirements were determined using existing Army job inventory procedures and by applying one or more judgmental approaches. For example, the critical incident technique was used to develop performance rating scales, and the Army Occupational Survey job inventory approach was used to help identify important MOS-specific tasks. The accuracy, completeness and appropriateness of the task and job information obtained was assessed in the development of the criterion measures. A detailed description of the task analytic and measures construction strategies appears in the Project A Annual Reports (1983, 1984, 1985) and previous Annual Reports to Congress (1983, 1984, 1985).

Pretesting Strategy

The completed hands-on test package was pilot tested with representative scorers and soldiers. The purpose of the field testing was to assure that the test could be administered as designed in a field environment and to determine scorer reliability. Subjective data on acceptability and feasibility were also collected from scorers and examinees. The pretesting strategies are described in greater detail under Field Testing Accomplishments.

Sampling Approach

The sampling plan specifies which MOS will be examined from the universe of possible MOS and details sample sizes for first-term enlisted personnel within each MOS. The selection of the 19 MOS was described previously. Detailed information regarding more extensive data collections planned for these and other measures is included in the research plan for "Improving the Selection, Classification and Utilization of Army Enlisted Personnel" (1983).

Data Collection Procedures

Data collection began with a briefing of local military commanders, examination of the test sites, equipment, and supplies, training of test administrators and scorers, and a dry run of testing procedures. Test site managers were appointed to supervise the actual data collections and were responsible for controlling the quality and flow of data from the testing.

Analyses to be Conducted

Statistical analyses focus on reliability, validity and test fairness issues for the training, job, and service-wide performance measures. Reliability is being assessed using (1) test-retest procedures, (2) variance partitioning procedures to estimate generalizability of test results over variables such as task type, scorer, test station, etc., and (3) interrater agreement estimates. Validation of selection and classification measures focuses on the content, construct, and concurrent validity of the battery with respect to training, job, and service-wide performance. Analysis-of-variance techniques were used to examine predictors for possible sex and race/ethnicity subgroup bias. Project analyses are discussed in full in the Project A Annual Reports (1983, 1984, 1985).

CURRENT STATUS AND ACCOMPLISHMENTS

A complete report of accomplishments and current status of the Army's research to link enlistment standards to job performance has been published by the Army Research Institute (1983, 1984, 1985) and is available for distribution. What follows includes a summary of accomplishments in the area of job performance measurement during 1986.

Performance Measures Development

Job knowledge tests, hands-on tests, and job performance and Army-wide rating instruments were developed in 1984 for MOS 13B, 64C, 71L, and 95B. A report was prepared describing the rationale for and procedures followed in the analysis and selection of relevant job tasks for the hands-on and job-specific knowledge tests for these MOS (HumRRO and ARI, 1984; Campbell and Harris, 1985). Following the recommendations in that report, tests and rating instruments were developed in 1985 for the remaining MOS.

Task Selection and Instrument Construction. A detailed, technical account of the procedures for task selection and instrument construction is documented in Campbell and Harris (1985) and the Project A Annual Report (1985). A brief account is presented below.

The general model and procedures for performance measure criterion development in Project A is as follows: The basic cycle of a comprehensive literature review, conceptual development, scale construction, pilot testing, scale revision, field testing and proponent (management) review was followed for each kind of criterion measure. The primary goals of criterion measurement in Project A were to: a) make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency, b) compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e. a multi-trait, multi-method approach), c) develop rating scale measures of performance factors that are common to all first tour enlisted MOS (Army-Wide measures), d) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and e) evaluate existing archival and administrative records as possible indicators of job performance.

For the hands-on measures, a comprehensive task sampling procedure was used to define the population of tasks in each MOS and then select 30 job tasks to represent the population of the MOS tasks. The task lists were then reviewed by the proponent schools for completeness and representativeness of the occupation. Fifteen tasks requiring a high level of physical skill, a series of prescribed steps, and speed of performance were selected for hands-on testing. The test items for the hands-on measures were generated from training manuals, field manuals, interviews with officers and job incumbents, and other appropriate sources.

The job knowledge tests, a paper and pencil multiple choice format, were developed to cover all of the thirty tasks in the MOS lists. The item content was generated on the basis of training materials, job analysis information, and interviews.

Two types of rating scales were also developed. One type of seven-point scale was designed to be parallel to the job tasks that were measured in the hands-on mode; one scale was developed for each of the fifteen tasks. The second type of rating scale followed standard procedures for developing Behaviorally Anchored

Rating Scales from "critical incident" workshops involving 70-75 officers and NCO's. This procedure resulted in six to nine MOS-specific Behaviorally Anchored Rating Scales, or BARS, for each of the nine MOS. A similar procedure was used also to develop Army-Wide (A-W) performance rating scales.

Training Knowledge (achievement) tests were also developed. These tests were based upon training course content. The content distribution of items on the test was proportional to the content of the course. The item pool was written by a team of subject matter experts, contracted for that purpose; the items were edited for clarity and relevance to training and job performance prior to field testing.

Field Testing Accomplishments

Field testing of the instruments for MOS 13B, 64C, 71L, and 95B was completed in 1984 and described in the Third Annual Report to Congress. The performance measurement instruments for MOS 11B, 19E, 31C, 63B, and 91A were field tested in 1985 and described in the Fourth Annual Report to Congress.

Sample Description. The field test MOS sample size by site data were provided in the Third Annual Report to Congress. A total of 1369 soldiers in the nine MOS were tested at six different sites. All soldiers were in Skill Level One, entry-level positions in their respective MOS and had entered the Army between 1 April 1982 and 30 June 1983.

Test Administration. Each test site had a test site manager who supervised all of the research activity and maintained the orderly flow of personnel through the data collection stations. An officer and two NCOs from one of the supporting units at the test site were assigned to support the field test. The officer provided liaison between the data collection team and the tested units; the NCOs coordinated the flow of equipment and personnel through the data collection procedures.

During the week preceding data collection at each research site the NCO scorers for the hands-on measure were given one day of training on scoring procedures by members of the research staff. The scorers were told about the overall design and nature of Project A; their critical influence on the reliability and validity of the measures was emphasized. Test administrators (contractor research staff and Army civilian research staff) for the remaining criterion measures were also trained at that time on the test administration procedures and one test administrator was assigned to each test station.

For data collection purposes, the criterion measures were divided into four major blocks:

1) Hands-on measures
2) Rating measures
3) Paper and pencil job knowledge tests
4) Paper and pencil training achievement tests.

Each block comprised one half day of participant time and each participant was tested for a two-day period. The order of administration of the blocks of measures was counterbalanced at each test site. Data collection at each site required approximately two weeks to complete.

Lessons learned from the field testing fall into three broad categories: logistics, scorer and test administrator training, and test site management. Based on the experience with the field tests, procedural modifications were implemented for the concurrent validation. Modifications include, but are not limited to: earlier coordination with Army support personnel for the test site; use of the Army-wide personnel locator to name-request soldiers for testing, rather than designating units to supply all their appropriate soldiers; intensive one-week training course for test administrators prior to arrival at test sites; development of a hands-on scorer certification program; and improved systems for conducting the orderly flow of personnel through data collection stations.

Reliability Results. Early reliability analyses of the field test data were used to provide information for the subsequent revision of the criterion measures to be used in the concurrent validation. Accordingly, measures were revised, if needed, to improve their statistical reliabilities and factor structures. This revision process included the elimination or modification of some test items; however, the revisions did not compromise the range of item difficulty (easy to hard) present in the measures. Following the revision and prior to the concurrent validation data collection, each measure was officially reviewed for content and accuracy and approved by the Commanding General with proponency for the respective MOS.

The results of the statistical analyses are presented in detail in the field test reports (1985) and the Project A Annual Report (1985).

Performance Measures and ASVAB Relationships. Analyses of the interrelationships among the different types of performance (criterion) data (from hands-on, written knowledge, and rating measures) resulted in moderate correlations. The results indicated that different aspects of the job performance domain were being measured by the different testing instruments, as intended. Very high correlations would have indicated that the instruments were all measuring the same aspect of the performance domain. Very low correlations, on the other hand, might have indicated that the data were primarily a reflection of the different measurement methods. There were no plans to examine the relationships among the performance measures and ASVAB during this stage of the project. More detailed description and discussion of the results is presented in the field test and annual reports, cited earlier. The relationships between the criterion scores and ASVAB, as well as continuing analyses among the performance measures, were examined closely using the data collected in the concurrent validation phase of the research project.

Project Demonstration Accomplishments

The concurrent validation data collection began in June and ended in November 1985. As reported in the Fourth Annual Report to Congress, data were collected at fourteen different sites. The data from 5200 soldiers, who were tested on all the project criterion measures, were entered into the research data base. Data from an additional 4000 soldiers, who were tested only with the measures of training success and service-wide performance were also entered. Analyses began as soon as all the data were collected and entered in the longitudinal research data base.

Summary of Results to Date. This section focuses on the group of nine MOS receiving the full set of the project criterion measures. The summary provides information on a sample of the measures administered in the concurrent validation. More detailed information will be available in the Project A Annual Report (1986).

The split-half reliability estimates for the hands-on, job knowledge, and school knowledge measures, presented below, indicate a high degree of internal consistency in the measures for each MOS.

Internal Consistency Reliability Estimates
for the Hands-On, Job Knowledge, and School Knowledge Tests

	Test		
MOS	Hands-on	Job Knowledge	School Knowledge
11B	.54 (682)	.89 (678)	.93 (684)
13B	.75 (612)	.85 (639)	.89 (640)
19E	.63 (474)	.89 (459)	.93 (485)
31C	.79 (341)	.86 (326)	.93 (349)
63B	.52 (569)	.87 (596)	.94 (612)
64C	.64 (640)	.85 (668)	.90 (669)
71L	.73 (494)	.82 (501)	.88 (493)
91A	.60 (496)	.89 (483)	.92 (479)
95B	.58 (665)	.84 (665)	.88 (674)

Note: The first entry in each cell is the Spearman-Brown corrected split-half reliability estimate. The second entry (in parentheses) is the sample size.

Similarly, the inter-rater reliability estimates for the Behaviorally Anchored Rating Scales (BARS) demonstrate moderate to strong reliabilities among raters. The fact that the reliability estimates for supervisor ratings are generally higher than the estimates for peer ratings may be attributed to the fact that supervisors have more experience rating individuals than non-supervisors.

Inter-rater Reliability Estimates for
Behaviorally Anchored Rating Scales
(N-Rater Reliability)

			Scale	
MOS	No. of Ratees	Average No. Ratings/Ratee	Average of All A-W BARS	A-W BARS of Overall Effectiveness
11B				
Supervisors	652	1.9	.70	.64
Peers	679	3.4	.68	.62
13B				
Supervisors	638	1.9	.61	.50
Peers	633	3.4	.63	.54
19E				
Supervisors	490	1.9	.65	.57
Peers	485	3.3	.62	.52
31C				
Supervisors	349	1.8	.64	.53
Peers	316	2.7	.60	.40
63B				
Supervisors	597	1.9	.72	.61
Peers	552	2.6	.57	.49
64C				
Supervisors	628	1.8	.66	.63
Peers	645	3.6	.62	.52
71L				
Supervisors	460	1.7	.72	.66
Peers	422	2.3	.41	.35
91A				
Supervisors	467	2.0	.71	.60
Peers	480	3.2	.62	.52
95B				
Supervisors	625	1.9	.61	.59
Peers	681	3.7	.64	.61

The intercorrelations among the criterion measures for each of the nine MOS are presented below. The correlation coefficients are all non-zero and, for the most part, in the moderate range. This pattern of results is similar to that found in the field test data. As mentioned earlier in connection with the field test results, these moderate correlations indicate that different aspects of the job performance domain are being measured by the different testing instruments.

Intercorrelations Between Project A Measures

		Measure						
Measure		1	2	3	4	5	6	7
	MOS=11B N=695							
1	Army-wide BARS Supervisors	1.00						
2	Army-wide BARS Peers	.59	1.00					
3	Overall Effectiveness Supervisors	.88	.54	1.00				
4	Overall Effectiveness Peers	.53	.86	.49	1.00			
5	School Knowledge Test	.29	.29	.31	.30	1.00		
6	Job Knowledge Test	.31	.30	.29	.31	.68	1.00	
7	Hands-On Test	.28	.22	.30	.23	.37	.47	1.00
	MOS=13B N=665							
1		1.00						
2		.48	1.00					
3		.84	.41	1.00				
4		.44	.81	.37	1.00			
5		.24	.23	.24	.22	1.00		
6		.24	.22	.23	.18	.69	1.00	
7		.26	.15	.26	.14	.42	.37	1.00
	MOS=19E N=502							
1		1.00						
2		.45	1.00					
3		.84	.39	1.00				
4		.40	.81	.36	1.00			
5		.18	.26	.22	.24	1.00		
6		.21	.23	.20	.21	.75	1.00	
7		.10	.16	.11	.19	.29	.43	1.00
	MOS=31C N=358							
1		1.00						
2		.39	1.00					
3		.83	.36	1.00				
4		.28	.80	.26	1.00			
5		.14	.10	.25	.08	1.00		
6		.20	.16	.24	.12	.68	1.00	
7		.25	.12	.25	.16	.50	.51	1.00

(continued, next page)

Intercorrelations Between Project A Measures (con't)

	Measures						
Measures	1	2	3	4	5	6	7
MOS=63B N=625							
1	1.00						
2	.49	1.00					
3	.86	.48	1.00				
4	.43	.80	.41	1.00			
5	.23	.19	.22	.16	1.00		
6	.17	.16	.19	.14	.73	1.00	
7	.16	.09	.16	.08	.33	.32	1.00
MOS=64C N=678							
1	1.00						
2	.50	1.00					
3	.86	.47	1.00				
4	.45	.83	.42	1.00			
5	.17	.14	.14	.15	1.00		
6	.17	.15	.13	.17	.65	1.00	
7	.18	.19	.16	.16	.36	.42	1.00
MOS=71L N=503							
1	1.00						
2	.40	1.00					
3	.86	.32	1.00				
4	.34	.78	.27	1.00			
5	.22	.13	.22	.16	1.00		
6	.20	.12	.18	.12	.71	1.00	
7	.13	.16	.16	.15	.58	.58	1.00
MOS=91A N=498							
1	1.00						
2	.51	1.00					
3	.87	.45	1.00				
4	.47	.82	.43	1.00			
5	.25	.22	.20	.24	1.00		
6	.20	.18	.19	.21	.65	1.00	
7	.18	.17	.17	.17	.46	.48	1.00
MOS=95B N=692							
1	1.00						
2	.52	1.00					
3	.85	.50	1.00				
4	.45	.85	.45	1.00			
5	.16	.19	.18	.16	1.00		
6	.12	.14	.15	.10	.58	1.00	
7	.18	.20	.21	.18	.30	.38	1.00

The Army's objectives for the criterion analyses were to identify an array of basic criterion scores, investigate the latent structure of these scores, and determine criterion construct scores. The objectives were accomplished through the application of a variety of statistical techniques. Basic criterion scores for rating measures were obtained by factor analysis of the individual scales to produce clusters of scales; the scales within each cluster were then averaged together. Basic scores for the hands-on and knowledge tests resulted from the clustering, by expert judgement sorts, of items into functional task clusters.

Exploratory factor analyses were then conducted on each MOS to produce hypotheses about the latent structure of the criterion space. The best-fitting model was tested using confirmatory factor analysis. The analyses resulted in the identification of five criterion constructs (factors):

1. Basic Soldiering Skills (use of basic weapons, first aid, etc.)

2. MOS Specific Technical Skills (document preparation for 71L; tank operation for 19E, etc.)

3. Exercise of Leadership, Effort and Self Development (the individual's willingness to perform the tasks and to be cooperative and supportive to other soldiers)

4. Maintaining Personal Discipline (adherence to Army regulations and traditions, committment to high standards of personal conduct)

5. Military Bearing/Fitness (maintenance of appropriate military appearance and good physical condition)

Once the scores comprising each criterion factor (construct) were identified, the scores were weighted and summed within each factor to obtain a construct score. The correlations between the construct scores and AFQT scores for the nine MOS are presented below. It should be noted that for 11B all skills are basic soldiering skills; there are no MOS-unique skills for the infantryman. Examination of the correlations indicates that AFQT predicts Basic skills and, to some extent, Technical Knowledge/Skill construct scores. The AFQT does not predict the other three construct scores; correlations between AFQT scores and Effort/Leadership, Personal Discipline and Appearance/Fitness scores are generally not statistically different from zero. These results are not unexpected since the AFQT ASVAB measure was developed to predict general cognitive abilities (trainability) for first term military enlistment (selection) purposes.

Correlation Between AFQT and Criterion Construct Scores

MOS	N	Basic Soldier Skills	Technical Knowledge/ Skill	Effort Leadership	Personal Discipline	Appearance/ Fitness
11B	478	.42 (.58)	- -	.19 (.40)	.15 (.25)	-.03 (.07)
13B	426	.34 (.43)	.23 (.33)	.11 (.29)	.05 (.13)	-.10 (.00)
19E	367	.46 (.49)	.26 (.23)	.17 (.25)	.13 (.08)	-.06 (-.02)
31C	274	.39 (.59)	.41 (.65)	.08 (.10)	.04 (.00)	-.16 (-.28)
63B	461	.31 (.48)	.24 (.51)	.04 (.25)	-.01 (.02)	-.04 (-.10)
64C	486	.37 (.60)	.23 (.43)	.00 (.10)	-.04 (-.14)	-.04 (-.10)
71L	407	.40 (.57)	.43 (.56)	.16 (.28)	.10 (.01)	.00 (.05)
91A	378	.31 (.68)	.34 (.67)	.09 (.21)	.07 (.05)	-.01 (-.24)
95B	569	.30 (.64)	.28 (.61)	.13 (.40)	.10 (.31)	-.01 (-.12)

Note: Numbers in parentheses are corrected correlation coefficients. The correction for range restriction employed the multivariate correction procedure designated by the Joint Services Job Performance Measurement Working Group (Working Group Minutes of the 9-10 July 1985 meeting, 26 August 1985)

As noted in the beginning of this report, the Army's goals include the development and validation of new and/or improved selection and classification measures. Preliminary analyses of the Army project's additional (e.g., non-cognitive) predictor measures indicate that the criterion construct scores poorly predicted by AFQT are well-predicted by other Project A measures. However, it is not the Army's intention to predict individually the construct scores. The Army will be developing an overall performance score, a weighted construct composite score, for each MOS. The weights will be obtained in workshops with NCOs and company grade officers from each MOS. They will scale the relative importance of each criterion construct for overall performance. The composite scores for each MOS will be computed and used to derive a single prediction equation for each MOS. Since this portion of the research has not been completed, results will be presented in future reports.

NATIONAL ACADEMY OF SCIENCES' ANNUAL REPORT

The Army appreciates the careful attention given to the Joint Service Project by the National Academy of Sciences. The Army's detailed responses to the issues raised by the Academy are presented below.

Army Effort to Implement National Academy of Sciences' Recommendations

In order to provide the appropriate context for the Army's response, it must be noted that the Army has completed the testing falling within the original Congressional and Department of Defense mandate to link enlistment standards to job performance. Measures of first tour enlisted job performance have been developed and concurrent validation data have been collected to examine the linkage of these measures to enlistment standards.

However, the Army's project is not yet complete; a longitudinal data collection is planned involving the administration of performance measures in 1988 and 1991. The comparison between the concurrent and longitudinal validation results has considerable importance for both the Army and scientific communities. Such a comparison would be jeopardized if the Army substantially alters its performance measures or testing procedures for the longitudinal validation.

Within the context of the constraints mentioned above, the Army will explore the feasibility of incorporating the Academy's recommendations to the extent possible in the longitudinal validation or through post hoc analyses of data already collected.

Construction of Performance Tests. The Academy's recommendations on performance test construction focus on three separate areas: competence measurement, job analysis, and task selection. The Army's response addresses each of the areas in turn.

A reasonable first step toward developing competency measures, and one which the Army is now considering, is to examine the performance measures already developed for the Joint Service Project, and examine the technical feasibility of determining with some degree of confidence and consensus what constitutes "mastery" on these measures. To the extent the Army is able to identify standards of job mastery, it will necessarily develop multiple scoring strategies.

With respect to job analysis procedures, the Army has carefully considered the recommendation that a more thorough-going job analysis be conducted than one that relies on task inventories and believes that it has addressed this recommendation by following a job analysis strategy which supplements the information provided by such inventories in several important ways. A "task" listed in an inventory did not provide the basis for Army performance measures. As the National Academy points out, task lists "tend to strip job information of its contextual setting, with the danger that the job is trivialized, its essence lost," and furthermore "the information given by task inventories tends to be very spare (Wigdor and Green, p. 42)." Army tasks were consolidated into meaningful functional groupings. In some instances, these groupings corresponded to "tasks" described in Army Soldier's Manuals. It should here be pointed out that a Soldier's Manual "task" bears no resemblance to a task found on an inventory. A Soldier's Manual task is more comprehensive. The Soldier's Manual task description incorporates the conditions ("contextual setting") under which the task is to be performed, "standards" or training objectives associated with the task, and performance measures, or steps to be accomplished in performing the task. A task inventory "task" might correspond to one of several steps which comprise a Soldier's Manual task.

Task inventory "tasks", or statements, provided information on what activities soldiers performed. Where statements did not correspond to elements of Soldier's Manual tasks, they were grouped to form new tasks. In order to ensure that statements were grouped in meaningful ways and placed in an appropriate "contextual setting", the Army relied on Technical Manuals, other supporting Army publications, subject matter expert input and direct task observation, where necessary.

The suggestion that the Services should provide information about the "behavioral content of tasks" is a somewhat more complicated one. There is the suggestion of need to "organize task statements into broader categories of behavior (Wigdor and Green, p. 42)." Then one can determine the underlying "measurable human attribute" and whether it requires "measurement in cognitive, affective, or psychomotor domains (p. 43)."

These statements suggest that behavioral content information is important to inform us about underlying attributes. We find this recommendation somewhat perplexing if it suggests that hands-on performance measures should be developed to represent underlying attributes rather than the behavioral steps required to perform a task. Such a developmental strategy would, in our view, lead us further from, rather than closer to, measures with demonstrable content validity. Perhaps we should instead interpret the recommendation to mean that we need information about underlying attributes as a basis for developing predictor tests. This would be a position that the Army would endorse, and is one which is indeed consistent with the procedures used in developing the Army projects new predictor measures.

Regarding the Academy's recommendations pertaining to task selection, the Army has carefully considered the recommendations. The major difficulty foreseen in implementing this recommendation is that the credibility of the task selection process to the policy makers who will ultimately determine how the findings from the Joint-Service Project will be implemented is not considered. The validity coefficients reported will not be credible if the performance measures are not credible. The performance measures will not be credible if they do not appear to provide adequate representation of important tasks. Thus, it is the Army's position that task importance should be a factor in task selection and that accordingly, random sampling techniques would not be appropriate.

The Army has focused considerable attention on standardizing and documenting the judgment-based procedures it used, as demonstrated by the summary in the Committee's annual report (p. 47) and in the Army report cited therein (Human Resources Research Organization and American Institutes for Research, 1984).

Selecting and Training Test Administrators. There are clear trade-offs involved in the decision of whether to use civilians or active-duty personnel as test administrators. On the one hand, a civilian administrator offers the potential advantages of standardization and "neutrality" in the scoring role. On the other hand, the "job expert" requirements for the test administrator create demands that will often be difficult for a civilian to meet. The test administrator must be familiar with the equipment used for reasons of both safety and scoring accuracy. Variations in equipment from post to post suggest that the best way to insure such familiarity is to use a scorer stationed at the post where the test is administered. Scorers must be familiar with local standard operating procedures which impact upon the correctness of a particular examinee response.

The credibility of the entire research effort depends to some extent on the technical competence of the scorers. Even a retired Army officer or NCO quickly loses currency regarding military policies and procedures and would face a potential "credibility gap" trying to evaluate soldiers on their performance. To the extent soldiers lose confidence in the expertise of the scorers, they will take the test less seriously. Active duty military personnel give the testing situation an air of authenticity and gravity it would not otherwise have.

The Army has chosen to use active-duty personnel as scorers. While having them avoid evaluation of examinees under their supervision is in some instances impossible, the Army has undertaken to determine how frequently such evaluation occurs and its impact on examinee scores. Analyses for one Army military occupational specialty, infantryman, indicate that the percent of soldiers scored by a supervisor on any particular task averaged less than 3%, and the correlation between the score received and the existence of a supervisory relationship between scorer and examinee ranged across tasks from .02 to .09, with a median correlational value of .03 (G. Hoffman & P. Ford, personal communication, August 8 1986). These results indicate that such a relationship has a minimal or negligible impact upon the examinee hands-on test score.

Within the limits associated with numbers of available qualified military personnel at a given post, the scientific Army staff has been actively involved in the selection of test administrators. Scorers were observed and evaluated both during scorer training and during hands-on testing and, when necessary, scorers were replaced.

The recommendation that a rater (scorer) training program be developed was followed by the Army. The program, although limited to one day, contains the major features recommended by the Committee. Test Administrators/Scorers were trained on scoring their tasks, were observed scoring one another on each relevant task, and were provided with appropriate feedback. The written instructions read to the scorers by the Hands-On Manager specifically emphasized "the need for standardized administrative procedures and the detrimental effects that any departure from the standardized procedures would produce." For example, note these instructions: "When you read the instructions, and in all your contact with soldiers being tested, your facial expression, voice inflection, and posture should be the same. You may hope people you like do well and people you do not like do poorly, but you must treat everyone the same. Your facial expression, voice, and posture must not threaten soldiers you test. Your demeanor should be objective, professional, and non-threatening" (Campbell, 1985). The instructions place particular emphasis on the need to avoid providing feedback to soldiers.

The rater training program was fully documented in hard copy and include Test Scorer Instructions and Scoresheet, sample test materials, Instructions for the Hands-On Manager, Hands-On Scorer Orientation Overview, Scorer Orientation Script, and Instructions for Scorer Certification. The materials, as recommended by the Committee report, provide information to administrators on "what to do, how to do it, and why it is done that way."

Comparing Performance Measures. The Army has developed and administered, for all nine military occupational specialties for which hands-on measures have been developed, ratings and written job knowledge tests as well. Having completed development and full-scale administration of first-tour measures, the Army has limited flexibility to modify developmental procedures to produce new kinds of performance measures. Nevertheless, the Army will consider the feasibility of following the Academy's recommendation for direct comparison of performance measures.

While the Army will have trouble implementing this recommendation in its own project at this point in time, the Army scientists endorse the spirit of the recommendation. Therefore, the Army project scientists have provided the Air Force scientists with materials to support the effort to develop Army project measures for use in testing an Air Force job. The Army will provide additional information to the Air Force as needed.

RESOURCES

The funding and in-house manpower resources for the Army's research to link enlistment standards to job performance are provided below.

Funding

Resources and manpower are not treated here as mutually exclusive categories. Estimated in-house manpower expenses are incorporated into the funding estimates.

Category	FY1987	FY1988	FY1989	FY1990	FY1991
6.2	1.0M	1.0M	.8M	.4M	
6.3A	3.5M	3.4M	2.3M	1.2M	1.1M

Manpower

PSY	FY1987	FY1988	FY1989	FY1990	FY1991
	7	7	7	7	7

REFERENCES

Campbell, J.P., and Harris, J.H. (1985, August). Criterion reduction and combination via a participative decision making panel. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles, CA.

Campbell, R.C. (1985). Scorer training materials (RS Working Paper 85): Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO), American Institutes for Research (AIR). (1984). Selecting job tasks for criterion tests of MOS proficiency. (RS-WP-84-25). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), Personnel Decisions Research Institute (PDRI), and Army Research Institute (ARI). (1983). Improving the Selection, Classification and Utilization of Army Enlisted Personnel: Annual Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), Personnel Decisions Research Institute (PDRI), and Army Research Institute (ARI). (1983). Improving the Selection, Classification and Utilization of Army Enlisted Personnel: Annual Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), Personnel Decisions Research Institute (PDRI), and Army Research Institute (ARI). (1984). Improving the Selection, Classification and Utilization of Army Enlisted Personnel: Annual Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), Personnel Decisions Research Institute (PDRI), and Army Research Institute (ARI). (1985, in press). Improving the Selection, Classification and Utilization of Army Enlisted Personnel: Annual Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), Personnel Decisions Research Institute (PDRI), and Army Research Institute (ARI). (1985, in press). Report on the Results of the Field Tests. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), Personnel Decisions Research Institute (PDRI), and Army Research Institute (ARI). (1986, in preparation). Improving the Selection, Classification and Utilization of Army Enlisted Personnel: Annual Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences

Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics). (1983). A Report to the House Committee on Appropriations: Second Annual Report to Congress on Joint-Service Efforts to Link Standards for Enlistment to On-the-Job Performance.

Office of the Assistant Secretary of Defense (Manpower, Installations, and logistics). (1984). A Report to the House Committee on Appropriations: Third Annual Report to Congress on Joint-Service Efforts to Link Enlistment Standards to Job Performance

Office of the Assistant Secretary of Defense (Force Management and Personnel). (1985). A Report to the House Committee on Appropriations: Fourth Annual Report to Congress on Joint-Service Efforts to Link Enlistment Standards to Job Performance.

Wigdor, A.K., and Green, B.F. (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project. Washington, D.C.: National Academy Press.

Wise, L.L., Wang, M., and Rossmeissl, P.G. (1983). Development and validation of Army selection and classification measures Project A: Longitudinal Research Database Plan. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

END
DATED
FILM
9—88
DTIC